UNIVERSITÉ CATHOLIQUE DE LOUVAIN
FACULTÉ DE PHILOSOPHIE, ARTS ET LETTRES

LES APPORTS DU TRAITEMENT AUTOMATIQUE DU LANGAGE À LA LISIBILITÉ DU FRANÇAIS LANGUE ÉTRANGÈRE

Thèse présentée par Thomas François en vue de l'obtention du titre de Docteur en Linguistique.

Jury : Pr Cédrick Fairon (promoteur)
Pr Anne Catherine Simon (co-promotrice)
Pr Georges Antoniadis
Pr Liesbeth Degand
Pr Piet Desmet
Pr Marco Saerens
Pr Heinz Bouillon (président du jury)



LOUVAIN-LA-NEUVE
ANNÉE ACADÉMIQUE 2011-2012
CENTRE DE TRAITEMENT AUTOMATIQUE DU LANGAGE

Table des matières

Introduction **27**

I Bases théoriques : lecture et lisibilité 29

1 La lecture en FLE 31

1.1 Introduction 31

1.2 Les processus de lecture en langue première et seconde 33

1.2.1 Essai de définition : la lecture 34

1.2.2 Les grandes problématiques de la lecture : une approche componentielle 42

1.2.3 Les théories de la lecture 58

1.3 La place de la lecture et du texte écrit dans les différentes approches didactiques en FLE 85

1.3.1 La méthodologie traditionnelle (MT) 85

1.3.2 La méthodologie directe (MD) 86

1.3.3 La méthodologie active (MA) 87

1.3.4 La méthodologie audio-orale (MAO) 88

1.3.5 La méthodologie audio-visuelle (MAV) 89

1.3.6 L'approche communicative 90

1.3.7 La perspective actionnelle 91

1.3.8 Conclusion 92

1.4 Conclusion 92

2 Une approche particulière de la lecture : la lisibilité 95

2.1 Définition 95

2.1.1 La lisibilité 95

2.1.2 Les formules de lisibilité 97

2.2 Historique 101

2.2.1 Les origines 101

2.2.2 Les premières formules (1923-1938) 103

2.2.3 La période classique (1938-1953) 112

2.2.4 Perfectionnements et spécialisations (1953 à 1965) 127

2.2.5 La révolution de la closure (1965 à 1979) 142

2.2.6 Critiques et élans structuro-cognitivistes (1979 à 1989) 163

2.2.7 L'après-cognitivisme : le pragmatisme désenchanté (1990 à 2003) . 178

2.2.8 La lisibilité computationelle (2004 à nos jours) 184

2.2.9 Conclusion sur les études en langue maternelle ou langue première (L1) : un parcours sinueux 194

2.3 La lisibilité pour les langues étrangères 196

2.3.1 Introduction à la lisibilité en langues étrangères 196

2.3.2 Les premiers travaux 197

2.3.3 La période classique 199

2.3.4 Les recherches récentes pour l'anglais langue seconde ou langue étrangère (L2) 200

2.3.5 Travaux récents en FLE : pas grand chose à se mettre sous la dent ! 204

2.3.6 Conclusion 206

2.4 Conclusion générale 207

II Méthodologie : un modèle pour la mesure de la difficulté 209

3 Quelles variables pour un modèle en FLE ? 215

3.1 Introduction 215

3.2 Quelles variables pour un modèle de lisibilité en FLE ? 216

3.2.1 Caractéristiques d'un bon prédicteur en lisibilité 216

3.2.2 Les variables au niveau lexical 218

3.2.3 Les variables au niveau syntaxique 245

3.2.4 Les variables de la difficulté conceptuelle 261

3.2.5 Variables de la cohérence textuelle 275

3.2.6 Variables spécifiques à la lecture en L2 285

3.3 La paramétrisation des variables retenues pour notre modèle 294

3.3.1 Les variables lexicales 294

3.3.2 Les variables syntaxiques 305

3.3.3 Les aspects sémantiques 308

3.3.4 Variables spécifiques aux langues étrangères 313

3.3.5 Conclusion 317

3.4 Conclusion générale 318

4 Corpus d'entraînement : comment mesurer la difficulté des textes ? 321

4.1 Introduction 321

4.2 Le choix du critère 323

4.2.1 Introduction : qu'est-ce qu'un critère ? 323

4.2.2 L'avis d'experts 323

4.2.3 Tests de compréhension 325

4.2.4 Le test de closure 326

4.2.5 La vitesse de lecture 332

4.2.6 Le rappel 334
4.2.7 Autres critères 334
4.2.8 Un choix pratique : les manuels scolaires 336
4.2.9 Conclusion 338
4.3 Le corpus et sa constitution 339
4.3.1 Introduction : nos besoins 339
4.3.2 Le Cadre européen commun de référence pour les langues 341
4.3.3 Les critères de sélection 343
4.3.4 Les manuels retenus et leur classication 345
4.3.5 Caractéristiques du corpus 351
4.3.6 Conclusion 355
4.4 Deux problèmes 356
4.4.1 L'homogénéité 356
4.4.2 Le nombre de textes par classe 365
4.4.3 Conclusion 369
4.5 Conclusion générale 370

III Expérimentations et résultats 373

5 Résultats 375

5.1 Introduction 375
5.2 Analyse de la capacité prédictive des variables explicatives 376
5.2.1 Introduction 376
5.2.2 Outils de mesure d'une bonne variable 376
5.2.3 Constitution des corpus de tests 382
5.2.4 Analyse des variables lexicales 388

5.2.5 Analyse des variables syntaxiques 405
5.2.6 Analyse des variables sémantiques 415
5.2.7 Analyse des variables spécifiques au FLE 422
5.2.8 Conclusion : quelle efficacité pour nos prédicteurs ? 427
5.3 Vers un modèle de lisibilité 428
5.3.1 Introduction 428
5.3.2 Nombre de variables dans une formule et stratégie de sélection . . . 429
5.3.3 La pondération des variables 431
5.4 Comparaison des modèles 443
5.4.1 Introduction 443
5.4.2 Méthodologie d'évaluation : les corpus 444
5.4.3 Méthodologie d'évaluation : les mesures 445
5.4.4 Comparaison des modèles experts 448
5.4.5 Comparaison des modèles par sélection automatisée de variables . . 459
5.4.6 La formule finale 466
5.5 Conclusion 471

Conclusion **482**

A Tables des sigles et acronymes **483**

B Liste alphabétique des indices de lisibilité utilisés **487**

C Tableau d'analyse des variables indépendantes **501**

D Résultats des procédures d'optimisation des méta-paramètres **511**

D.1 La « baseline » 511
D.1.1 Les arbres de décision 511

D.1.2 Le bagging 512
D.1.3 Le boosting 512
D.1.4 Les SVM 514
D.2 Les modèles *Expert1* 518
D.2.1 Les arbres de décision 518
D.2.2 Le bagging 519
D.2.3 Le boosting 519
D.2.4 Les SVM 521
D.3 Les modèles *Expert2* 525
D.3.1 Les arbres de décision 525
D.3.2 Le bagging 526
D.3.3 Le boosting 526
D.3.4 Les SVM 527
D.4 Les modèles par sélection automatique de variables 531
D.4.1 Les arbres de décision 531
D.4.2 Le bagging 532
D.4.3 Le boosting 532
D.4.4 Les SVM 534

Table des figures

1.1 Modèle de Gough (1972) 63

1.2 Modèle de Rumelhart (1977) 67

1.3 Modèle de Just and Carpenter (1980) 70

1.4 Modèle de McClelland and Rumelhart (1981) : les voisins de la lettre T. . . 73

1.5 Modèle de Seidenberg and McClelland (1989). Seules les parties en gras ont été implémentées. 75

1.6 Le modèle BIA (Bilingual Interactive Activation) de van Heuven, Dijkstra et Grainger (1998). 77

1.7 Modèle DRC de Coltheart, Rastle, Perry, Langdon and Ziegler (2001). . . . 79

3.1 Image tirée de Wikipédia : http ://en.wikipedia.org/wiki/Latent_semantic_analysis 312

5.1 Coefficient de corrélation entre deux variables en relation linéaire 380

5.2 Coefficient de corrélation entre deux variables en relation non linéaire . . . 380

5.3 Nuage de points entre X75FLC et la difficulté des textes 390

5.4 Variations du coefficient de corrélation de Spearman pour **NormTTR_L** en fonction de n. 400

5.5 Nuage de points entre NMP et la difficulté des textes 406

5.6 Nuage de points entre NMP et la difficulté des textes 409

5.7 Valeurs de ρ en fonction de k pour **avLocalLsa_Tok** et **avLocalLsa_Lem**, calculées sur le *Corpus6CE*. 420

5.8 Évolution de la proportion moyenne des structures NA, NPN et NN par niveau de difficulté. Obtenu à partir d'un des dix échantillons de *Corpus6CE* ayant servi à calculer les corrélations. 424

5.9 Exemple de fonction pour $E(Y = 1 \mid X = x)$, où X est la variable **NMP**. $Y = 0$ correspond aux textes de niveau A1, tandis que $Y = 1$ représentent ceux du niveau C2. 435

5.10 Exemple de deux hyperplans séparateurs. Le premier B1, maximise la marge comprise entre b11 et b12, tandis que B2 sépare correctement les données, mais possède une marge plus réduite. 442

5.11 Exemple de graphique présentant les résultats de la fonction objective (ici, R) en fonction du paramètre nbagg dans le bagging appliqué au *Dev6CE*. . 450

5.12 Différences entre le modèle *Expert2* complet et les modèles excluant l'une des variables, calculées par classifieur séparateur à vaste marge (SVM) sur *Test6CE*. Plus la différence est positive, plus la variable associée apporte une information nouvelle au modèle. 457

5.13 Différences entre le modèle *Expert2* complet et les modèles excluant l'une des variables, calculées par classifieur SVM sur *Test9CE*. Plus la différence est positive, plus la variable associée apporte une information nouvelle au modèle. 458

5.14 Interface de recherche de Dmesure. 480

5.15 Exemple de résultats de Dmesure pour la requête « maison » envoyée à un site de français langue étrangère (FLE). 481

5.16 Exemple d'un historique d'utilisateur dans Dmesure. 481

Liste des tableaux

3.1 Exemples de deux candidats dont le ratio de la log-vraisemblance n'est pas significatif sur un petit corpus et qui le devient à l'extrême sur un très large corpus. Ils sont comparés à un troisième candidat qui est une véritable collocation. 315

4.1 Nombre de textes par niveau, pour chaque série de manuels pris en compte . 350

4.2 Nombre de textes par genre et par niveau, ainsi que le nombre de tokens (entre parenthèses). 354

4.3 Proportion de chaque genre de textes par niveau. Le pourcentage entre parenthèses représente la proportion de ces mêmes genres par niveau en termes de nombre de tokens. 355

4.4 Nombre moyen de lettres par mots pour chaque manuel. Nombre moyen de mots par phrase pour chaque manuel. Les ouvrages présentant un problème de cohérence sont indiqués en gras. Les chiffres entre parenthèses réfèrent au numéro du manuel au sein de sa série. Comme on peut le noter, certains niveaux contiennent des textes issus de deux ouvrages différents de la même série. 358

4.5 Nombre moyen de mots par phrase pour chaque manuel. Les ouvrages présentant un problème de cohérence sont indiqués en gras. 359

4.6 Nombre moyen d'idées par texte pour chaque manuel. Les ouvrages présentant un problème de cohérence sont indiqués en gras. 360

4.7 Nombre moyen de lettres par mots pour chaque manuel. Les ouvrages présentant un problème de cohérence sont indiqués en gras. Les chiffres entre parenthèses réfèrent au numéro du manuel au sein de sa série. Comme on peut le noter, certains niveaux contiennent des textes issus de deux ouvrages différents de la même série. 361

4.8 Nombre moyen de mots par phrase pour chaque manuel. Les ouvrages présentant un problème de cohérence sont indiqués en gras. 361

4.9 P-valeur pour chacun des tests analyse de la variance (ANOVA) effectués sur les deux corpus. Un résultat inférieur à $0,05$ signifie que l'hypothèse d'homogénéité a été rejetée pour ce niveau en particulier. Les niveaux de significativité sont représentés comme suit : $p < 0,001 : ***$; $p < 0,01 : **$ et $p < 0,05 : *$. 364

4.10 P-valeur pour chacun des tests ANOVA effectués sur le nouveau Corpus6Cleaned. Un résultat inférieur à $0,05$ signifie que l'hypothèse d'homogénéité a été rejetée pour ce niveau en particulier. Les niveaux de significativité sont représentés comme suit : $p < 0,001 : ***$; $p < 0,01 : **$ et $p < 0,05 : *$. 365

4.11 Pour chacun des deux modèles, sont présentés le R^2 estimé sur le corpus d'entraînement, à partir d'une procédure de bootstrap .632 et sur notre échantillon de test. 367

4.12 Matrice de confusion pour le modèle entraîné sur Corpus6Equi 368

4.13 Matrice de confusion pour le modèle entraîné sur Corpus6Apriori 368

5.1 Variables de fréquences lexicales (lemmes) et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : $^1 : < 0.05$; $^2 : < 0.01$; $^3 : < 0.001$. 389

5.2 Variables de fréquences lexicales (formes fléchies) et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : $^1 : < 0.05$; $^2 : < 0.01$; $^3 : < 0.001$. 391

5.3 Variables de fréquences lexicales (formes fléchies désambiguïsées) et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : $^1 : < 0.05$; $^2 : < 0.01$; $^3 : < 0.001$. 392

5.4 Pourcentage d'absents et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : $^1 : < 0.05$; $^2 : < 0.01$; $^3 : < 0.001$. 394

5.5 Modèles unigrammes et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : $^1 : < 0.05$; $^2 : < 0.01$; $^3 : < 0.001$. 395

5.6 Modèles unigrammes et difficulté : résultats de l'expérience avec retrait des « outliers » dans les trois variables. 396

5.7 Variables de n-grammes et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : $^1 : < 0.05$; $^2 : < 0.01$; $^3 : < 0.001$. 397

5.8 Corrélations croisées (ρ de Spearman) entre les quatre meilleures variables n-grammes et quelques variables d'intérêt, calculées sur l'un des échantillons issus de *Corpus6CE*. 398

5.9 Mesures de la diversité lexicale et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : $^1 : < 0.05$; $^2 : < 0.01$; $^3 : < 0.001$. 399

5.10 Moyennes du Type Token Ratio (TTR) par niveau de difficulté, calculées sur un échantillon du **Corpus6CE**. 401

5.11 Mesures du voisinage orthographique et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : $^1 : < 0.05$; $^2 : < 0.01$; $^3 : < 0.001$. 401

5.12 Mesures du voisinage orthographique et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : $^1 : < 0.05$; $^2 : < 0.01$; $^3 : < 0.001$. 403

5.13 Variables de la longueur des phrases et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : $^1 : < 0.05$; $^2 : < 0.01$; $^3 : < 0.001$. 407

5.14 Variables basées sur les catégories grammaticales et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : $^1 : < 0.05$; $^2 : < 0.01$; $^3 : < 0.001$. 410

5.15 Corrélations croisées (ρ de Spearman) entre quelques variables catégorielles et les principaux indices lexicaux et syntaxiques, calculées sur l'un des 10 échantillons issus de *Corpus6CE*. 411

5.16 Caractéristiques verbales et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : $^1 : < 0.05$; $^2 : < 0.01$; $^3 : < 0.001$. 413

5.17 Corrélations croisées (ρ de Spearman) entre les meilleures variables verbales et trois indices lexicaux et syntaxiques typiques, calculées sur l'un des 10 échantillons issus de *Corpus6CE*, choisi au hasard. 414

5.18 Caractéristiques verbales et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : $^1 : < 0.05$; $^2 : < 0.01$; $^3 : < 0.001$. 416

5.19 Corrélations croisées (ρ de Spearman) entre quelques variables du taux de personnalisation textuel et **PRO/ALL**, calculées sur l'un des 10 échantillons issus de *Corpus6CE*. 417

5.20 Densité des idées et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : $^1 : < 0.05$; $^2 : < 0.01$; $^3 : < 0.001$. 418

5.21 Corrélations croisées (ρ de Spearman) entre **ConcDens** et quelques variables typiques, calculées sur l'un des 10 échantillons issus de *Corpus6CE*. 418

5.22 Taux de cohérence et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : $^1 : < 0.05$; $^2 : < 0.01$; $^3 : < 0.001$. 421

5.23 Variables d'unités polylexicales (UPs) et difficulté : ρ de Spearman pour les deux corpus et les 4 seuils. Le taux de significativité des corrélations est indiqué comme suit : $^1 : < 0.05$; $^2 : < 0.01$; $^3 : < 0.001$. 423

5.24 Indices du type de texte et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : $^1 : < 0.05$; $^2 : < 0.01$; $^3 : < 0.001$. 426

5.25 Corrélations croisées (ρ de Spearman) entre **BINGUI** et quelques variables typiques, calculées sur l'un des 10 échantillons issus de *Corpus6CE*, sélectionné au hasard. 426

5.26 Classification des modèles statistiques utilisés dans cette étude d'après l'échelle de mesure des variables auxquelles ils s'appliquent. 433

5.27 Exemple de matrice de confusion pour un problème de classification binaire. Chaque entrée f_{ij} correspond au nombre d'observations de la classe i prédits comme appartenant à la classe j. 445

5.28 Performances des modèles par arbre de décision en fonction des paramètres cp et $minsplit$, estimées par validation croisée à dix échantillons sur *Dev6CE*.451

5.29 Performances des différents modèles « baseline » sur *Test6CE* et *Test9CE*. Les méta-paramètres utilisés dans leur développement sont également repris. 452

5.30 Performances des différents modèles Expert1 sur *Test6CE* et *Test9CE*. Les méta-paramètres utilisés dans leur développement sont également repris. . . 453

5.31 Performances des différents modèles Expert2 sur *Test6CE* et *Test9CE*. Les méta-paramètres utilisés dans leur développement sont également repris. . . 455

5.32 Corrélations croissées entre les variables de l'ensemble *Expert2*, calculées sur *Test6CE*. 457

5.33 Performances des différents modèles automatiques sur *Test6CE* et *Test9CE*. Les méta-paramètres utilisés dans leur développement sont également repris. 464

5.34 Détails sur les deux modèles finalement retenus pour les échelles à 6 et à 9 classes. 467

5.35 Exemple de matrice de confusion pour le modèle SVM à 6 classes. Elle correspond à une acc de 60% et une $acc-cont$ de 79%. Les comptes associés à des observations correctes, sous l'hypothèse de l'exactitude contiguë, sont graisés, tandis que les erreurs graves sont indiquées en rouge. 467

5.36 Taux d'exactitude contiguë par niveau pour le modèle SVM à 6 classes, calculée sur l'un des 10 échantillons de test dont l'$acc = 60\%$ et l'$acc - cont = 79\%$. 468

5.37 Exemple de matrice de confusion pour le modèle SVM à 9 classes. Elle correspond à une acc de 37% et une $acc-cont$ de $67,8\%$. Les comptes associés à des observations correctes, sous l'hypothèse de l'exactitude contiguë, sont graisés, tandis que les erreurs graves sont indiquées en rouge. 468

5.38 Taux d'exactitude contiguë par niveau pour le modèle SVM à 6 classes, calculée sur l'un des 10 échantillons de test dont l'$acc = 60\%$ et l'$acc - cont = 79\%$. 469

5.39 Comparaison des performances de notre modèle par rapport à celles de la littérature. 470

C.2 Variables d'UPs et difficulté : ρ de Spearman pour les deux corpus et les 4 seuils. Le taux de significativité des corrélations est indiqué comme suit : [1] : < 0.05 ; [2] : < 0.01 ; [3] : < 0.001. 509

Remerciements

> *À l'heure où commencent à se gercer les doigts roses de l'aurore, je montai tel un dard rapide dans un autobus à la puissante stature et aux yeux de vache de la ligne S au trajet sinueux (Queneau, « Exercices de style »).*

Émotif

Les remerciements, lorsqu'ils clôturent un travail aussi prenant, constituent toujours un moment particulièrement intense. Une larme perle subrepticement au coin de mon oeil, alors que je me remémorre la foule de bons moments vécus avec mes collègues et mes amis durant cette thèse. Il y a tout d'abord Laure Cuignet, qui a épanché nombre de chagrins lorsque la dernière amélioration de mon modèle ne me permettait pas de gagner plus d'un demi pourcent d'exactitude. Il y a ensuite tous les membres du kot Manga, Ookami, Olivier, Max, Rim, Cosma, Ephy, François, Mélanie, Gouzy, Jink, Grégory, Frère, Élise, Cid, Caroline, Marie, Simon, Gus, Ouki et Jean-Pierre, si chers à mon coeur, malgré qu'ils n'ont cessé de me répéter que pour bien réussir une thèse, il faut commencer par bien se détendre... Et il faut reconnaître qu'ils ont parfaitement réussi à me faire oublier, par moment, l'effrayante charge de travail qui m'attendait tapie au coin de mon ordinateur.

Il faut dire qu'ils ont reçu, dans cette mission, l'aide inespérée de plusieurs de mes collègues attentionnés. En effet, que serais-je sans les nombreuses fins d'après-midi au Gin et Juice où tant Amélie et Patrick que Noémi et Jenny, ou encore Deniz ou Jean-Léon, se sont parfois délestés de quantité de ferrailles pour me relancer de plus belle à l'ouvrage ? Enhardi par leur chaleur humaine et d'aimables breuvages diversement alcoolisés, je pouvais alors rentrer chez moi et y pourfendre les résistances méthodologiques qui se dressaient sur mon chemin. C'est encore eux qui m'ont consolé, lorsque tel Villon, je me larmoyais en disant : « Bien sais, se j'eusse moins estudié, j'eusse boisson et couche folle ».

Entre deux larmes, plus franches cette fois, je vous répète encore, mes chers amis : soyez

remerciés pour votre affection et vos bienfaits. Je les conserve au creux de mon coeur.

Dithyrambique

Les remerciements, lorsqu'ils clôturent un travail aussi imposant, constituent toujours un moment extrêmement particulier. Ils offrent l'occasion de rendre enfin justice à la pléthore de professeurs concernés et de collègues dévoués sans qui l'objectif n'aurait sans doute pas été atteint. Il va de soit que les premières personnes à qui je pense ici sont mes deux promoteurs, Cédrick Fairon et Anne Catherine Simon, dont les conseils avisés ont éclairé le chemin brumeux des premiers mois de ma recherche, avant de rehausser ensuite les sentiers de traverse empruntés au hasard de mes explorations de leurs avis critiques.

Qu'aurais-je été aussi sans le retour sagace des membres du comité qui a encadré mes premiers pas en lisibilité ? Jamais, je ne pourrais suffisamment exprimer l'étendue du respect mêlé d'une vive gratitude que j'éprouve à l'égard de Marco Saerens et Serge Verlinde. Toutefois, cette reconnaissance est encore éclipsée par le plaisir éprouvé lors de la discussion de ma défense privée pour lequel je remercie éminemment les différents membres de mon jury. Enfin, je tiens à souligner l'exceptionnel travail de Jean-Léon qui, guide sévère, mais bienveillant, a, tel Sisyphe, désherbé les larges plaines où mon esprit verbeux s'est étalé, aidé, dans cette tâche ingrate, par Hubert. C'est grâce à l'aide de ces érudits qu'il a été possible de remplir la noble mission que je m'étais fixée.

Minutieux

Les remerciements sont une chose sacrée, qui doit être faite dans les règles afin de n'oublier personne. Je suis ainsi remonté jusqu'au début de ma thèse, m'arrêtant sur le moment précis où tout a commencé. Le mercredi 30 août 2006 à 15h37, Cédrick Fairon m'a proposé de me défendre auprès du FNRS afin de réaliser cette thèse, ce dont je lui suis toujours particulièrement reconnaissant.

Néanmoins, j'ai eu vite des remords de me limiter à ce moment. Tout avait-il vraiment commencé ce jour-là ? Je me suis dis que si j'en étais arrivé à ce point, c'était aussi grâce à une série de personnes qui m'avaient formé à être capable de conduire une thèse à bien. Parmi celles-ci, les premiers sont certainement mes parents, Françoise et Marcel, qui m'ont poussé à être curieux de tout, ce qui a certainement joué un rôle dans ce choix de rédiger une thèse. Toutefois, malgré leur grande patience, ils n'ont pas pu résister à l'envie de confier épisodiquement leur enfant turbulent à de malheureuses victimes, contraintes de lui enseigner les bases de la vie en société et des rudiments de connaissances. Il me faut donc remercier la kyrielle de personnes ayant pris soin de m'éduquer durant ces longues années

qui précèdent la thèse : ma nounou, mon institutrice maternelle dont j'avais si peur, mon professeur de première année... Bon, je vois que je vous endors. Mettons qu'ils se reconnaîtront, cela m'évitera de perdre mon lecteur avant même le début de la thèse. Allez ! Je ne résiste pas à l'envie d'un dernier petit remerciement, à tous mes collègues, parmi lesquels je n'ai pas encore cité Bernadette, Richard, Adrien, Olga, Olivier, Nicolas, Claude, Sacha, Nadja, Kevin, Isabelle, Julia et Laurent, qui m'ont aidé à mater Linux, à me dépêtrer de mes labyrinthes algorithmiques en Python, Perl, Java, ou R, à élucider des notions de linguistique hermétiques, à m'épargner l'ennui des conférences à l'étranger, à numériser mon corpus, à programmer Dmesure, à rédiger des articles, etc.

Pressé

Devant boucler ces remerciements au plus vite avant impression, je remercie en vitesse toutes les personnes qui m'ont aidé, de près ou de loin, à réaliser ce travail et que j'aurais omises ailleurs dans ces remerciements. Pas le temps d'en dire plus. Si vous voulez plus de détails, reportez-vous à Minutieux.

Gourmand

Pour conclure ces remerciements, il ne faut pas oublier qu'un travail aussi consommateur en énergie n'aurait pas pu aboutir sans l'assistance culinaire de nombreuses personnes. Il convient ici de remercier à nouveau Laure, dont les petits plats mitonnés avec amour m'ont rendu le moral à plusieurs reprises, à défaut, lorsque leur consommation avait été excessive, de me relancer au travail avec une ardeur nouvelle. De même, je n'aurais jamais réussi à écrire un ouvrage à la tranche aussi épaisse sans les nombreux biscuits, gâteaux et autres délices offerts généreusement par mes collègues, en particulier Babette, Noémi, Stéphanie Weiser et Stéphanie Audrit.

Je me dois aussi de mentionner Patrick et Amélie pour leur recette de la courgette en dés accompagnée de boudin blanc et de purée. Celle-ci a avantageusement remplacé le boudin à la purée et compote de pommes que j'avais développé comme réflexe de survie lorsque Laure n'était pas là. Dans la même veine, je remercie Hubert pour son magnifique couteau en céramique, qui m'a permis de couper deux fois plus rapidement les poulets récalcitrants et d'ainsi me remettre plus vite à l'ouvrage. Enfin, il ne faut pas oublier nos quatre cordons bleus, Stéphanie Weiser, Stéphanie Audrit, Amélie et Sophie qui ont réconforté mon courageux auditoire au terme de la défense publique. Sans toutes ces personnes pour m'aider, je serais certainement mort d'inanition avant de taper le point final de cette thèse.

Introduction

Ce travail traite de la question des apports du Traitement Automatique du Langage (TAL) à la lisibilité du FLE. Il s'agit là d'une problématique spécifique au milieu de l'éducation et de l'enseignement des langues étrangères, dont l'intérêt peut être délicat à percevoir au premier abord. Afin de mieux faire comprendre celui-ci et avant d'expliquer ce qu'est la lisibilité, il convient de la replacer dans son contexte plus général : la lecture.

Depuis l'invention de l'écriture, son pendant, la lecture occupe une place prépondérante dans les sociétés humaines. En effet, elle a offert de merveilleuses possibilités à l'homme, dont celle de réceptionner la parole de scripteurs situés en amont du courant temporel, après que ceux-ci l'ont libérée de l'immédiat en la confiant à un support plus ou moins pérenne. Dès lors, il n'est pas étonnant que la lecture soit d'abord principalement associée à la transmission des écritures saintes au sein de ce que l'on appelle les religions du Livre (Judaïsme, Christianisme et Islam), ce qui lui confère une certaine dimension sacrée. Ainsi, elle n'est longtemps réservée qu'à une élite de clercs ou savants, tandis que la masse n'a accès au contenu des livres que via la lecture orale.

Un premier coup est porté à ce monopole avec la Réforme protestante, qui défend le droit de tous à lire et, surtout, à interpréter la Bible. Dans ce contexte qui oppose protestants et catholiques, « l'enseignement de la lecture devient un enjeu politique et religieux fondamental, réalisé via la création d'écoles. » (Horellou-Lafarge et Segré, 2003, 27). Par la suite, avec l'industrialisation croissante de la société, la nécessité de comprendre actes juridiques, contrats ou comptes touche une partie croissante de la population et conduit au développement d'une vision plus utilitariste de la lecture. La rationalisation s'empare de cette activité jadis sacrée et en fait un vecteur du savoir. Dès lors, dans les démocraties naissantes, ce sont les gouvernements, et non plus l'Église, qui organisent l'alphabétisation de l'ensemble de la population afin de permettre l'intégration harmonieuse de chacun au sein de la société industrielle.

Par conséquent, le livre, de réceptacle d'une parole institutionnelle, se diversifie, servant tour à tour à se distraire, à s'instruire, à se dépayser... En gagnant la capacité de lire, l'homme du commun acquiert du même coup l'opportunité d'entrer en contact avec un groupe beaucoup

plus large et avec des connaissances toujours plus diversifiées. Il devient capable de recul vis-à-vis de lui-même, mais aussi vis-à-vis de la société qui l'entoure.

Ainsi, l'aptitude à lire ne se limite pas à représenter un atout social pour celui qui la possède. Plus profondément que cela, il s'agit d'un mode de communication élargi qui permet à l'homme de transcender le temps et d'interroger sa propre époque et sa propre existence. La lecture joue donc un rôle essentiel aussi bien dans la constitution de la personnalité du lecteur que dans l'évolution des sociétés ; et cet aspect est d'autant plus pertinent qu'avec l'apparition d'Internet, l'homme a désormais accès à un espace de communication globalisé où il peut prendre connaissance de la pensée d'autrui hors d'un quelconque contrôle collectif.

On comprend pourquoi l'enseignement de la lecture en langue maternelle a toujours occupé une position centrale au sein de l'éducation : non seulement il constitue une étape préliminaire avant d'aborder d'autres domaines du savoir, mais il forme aussi des citoyens capables de susciter des changements sociaux suite à ce contact avec l'autre. À ce propos, il n'est pas étonnant de noter que le degré d'alphabétisation est également fortement corrélé avec le statut socioéconomique (Willms, 2003).

Toutefois, il subsiste une barrière à cette communication mondiale : la diversité des langues. Les outils de traduction automatique n'autorisant pas encore l'homme moderne à s'affranchir de cette limite, il lui faut continuer à s'efforcer d'apprendre à lire dans d'autres langues. Pourtant, plus que jamais confronté à des défis globaux, il gagne à communiquer avec les membres de communautés différentes, mais doit parfois y renoncer suite à ce constat que sa langue n'est pas universelle et qu'il ne peut apprendre toutes les langues. C'est notamment pour cette raison que le domaine de l'enseignement des langues étrangères, aujourd'hui si florissant, est devenu une préoccupation de l'Union européenne (UE).

Dans ce contexte, l'une des compétences que le professeur de langues étrangères cherche à transmettre à ses apprenants est la capacité à comprendre l'écrit. Il y a évidemment bien des approches de la question (qui sont discutées brièvement dans le Chapitre 1), mais il est aujourd'hui admis que cette mission passe par l'organisation d'exercices de lecture, qu'ils soient guidés ou indépendants. En effet, diverses expérimentations ont montré que la pratique régulière de la lecture permet d'améliorer ses compétences à ce niveau (Mastropieri *et al.*, 1999; Reitsma, 1988; Welsch, 2006). De même, lire dans une langue étrangère favoriserait plus globalement l'acquisition de celle-ci (Elley et Mangubhai, 1983; Iwahori, 2008; Krashen, 1989), particulièrement au niveau du vocabulaire (Pigada et Schmitt, 2006). Pour que cette pratique soit profitable, encore faut-il que les textes soient adaptés au niveau des étudiants (Bell, 2001; O'Connor *et al.*, 2002), ce qui n'est pas toujours le cas, d'après Mesnager (1986a), du moins en ce qui concerne la littérature enfantine francophone.

Malheureusement, il n'est pas si aisé pour un professeur de trouver des textes dont la difficulté correspond au niveau de compétence de ses étudiants, et qui sont susceptibles de les intéresser et de les motiver. Bien sûr, celui-ci peut se reposer sur une large gamme de manuels où la difficulté des matériaux est balisée. Toutefois, pour peu qu'il désire s'éloigner des sentiers battus en proposant un texte sur un sujet d'actualité ou une thématique qui lui tient à cœur, la recherche d'un document adapté au niveau de sa classe s'apparente parfois à un chemin de croix.

Cette problématique n'est pas neuve. Déjà dans les années 1920, des chercheurs et des éducateurs américains étaient désireux de mettre au point des techniques reproductibles qui soient capables d'associer un texte à un lecteur, quel que soit l'individu qui les mette en œuvre. C'est ainsi qu'est apparue la « lisibilité », un domaine dont l'objectif, très pragmatique, est d'évaluer la difficulté de textes pour une population donnée à l'aide d'une formule mathématique basée sur un nombre restreint de caractéristiques linguistiques. La conception de formules de lisibilité a rencontré un vif succès dans la culture anglo-saxonne, en particulier aux États-Unis, mais n'a guère réussi à pénétrer la sphère francophone (Timbal-Duclaux, 1985). De manière générale, il a connu une fortune changeante au cours du 20e siècle, avant d'être presque oublié durant les années 90.

Aujourd'hui, la lisibilité revient à la mode. Depuis 2004, on compte près d'une trentaine de publications concernant la lisibilité de l'anglais et, dans une moindre mesure, d'autres langues telles que le portugais ou l'allemand. Ce renouveau d'intérêt s'explique en grande partie grâce aux apports d'une autre discipline : le TAL. Celle-ci est apparue après la Seconde Guerre mondiale et vise à rendre les ordinateurs capables d'effectuer diverses tâches intégrant le langage, telles que permettre la communication homme-machine, faciliter la communication entre humains ou, plus simplement, traiter efficacement des données textuelles ou orales (Jurafsky et Martin, 2009, 35). Elle mêle ainsi diverses disciplines telles que la linguistique, l'informatique ou la statistique.

Cela fait déjà plusieurs années qu'on s'est rendu compte que le TAL pouvait aussi conduire à des avancées dans le domaine de l'enseignement des langues étrangères, plus particulièrement au sein de l'apprentissage des langues assisté par ordinateur (ALAO) (Tschichold, 2006). En effet, sa capacité à traiter, avec plus ou moins de bonheur, des phénomènes linguistiques complexes représente un atout lorsqu'il s'agit d'automatiser certaines tâches d'apprentissage d'une langue.

De la même façon, il semblerait que les possibilités offertes par le TAL aient inspiré les recherches récentes en lisibilité, du moins dans le monde anglo-saxon. En effet, comme lors de la première vague de travaux en lisibilité, le français n'est à nouveau guère touché par cette vague d'enthousiasme, que ce soit pour le français L1 ou le FLE. On ne recense

aucun travail qui se soit efforcé d'établir un pont entre ces deux disciplines : la lisibilité et le TAL [1]. Par ailleurs, les travaux récents sur le français L1 ou L2 sont tout simplement extrêmement rares (Barbaresi, 2011; Uitdenbogerd, 2005). Pourtant, les possibilités offertes par une formule de lisibilité en FLE sont loin d'être anecdotiques, comme le laissent penser les travaux sur l'anglais L1 et L2.

Sans prétendre être exhaustif, soulignons qu'une telle formule pourrait avoir des répercussions sur la pratique de la lecture dans un contexte éducatif, mais également être incluse dans des logiciels d'ALAO afin de rendre ceux-ci plus adaptatifs. Ainsi, en ce qui concerne la première de ces perspectives d'application, divers travaux (Bell, 2001; Iwahori, 2008) ont confirmé que la pratique extensive de la lecture – c'est-à-dire une approche qui prône la consultation d'un large nombre de textes dans un contexte de délassement – conduit les apprenants testés à faire des progrès plus importants que ceux auxquels est enseignée une pratique intensive de la lecture, où un texte est décortiqué sous toutes ses coutures au cours d'une leçon. Toutefois, pour que ce type de lecture extensive soit profitable, il convient, comme nous l'avons déjà dit, que les textes proposés aux apprenants soient bien de leur niveau, sous peine de les pousser à retomber dans une approche intensive, voire de les rebuter. Dès lors, l'acquisition automatique de textes sur l'internet est devenue récemment un souci partagé par plusieurs chercheurs issus de communautés linguistiques diverses (Collins-Thompson et Callan, 2004b; Miltsakaki, 2009; Uitdenbogerd, 2010). Ceux-ci ont développé des plateformes web permettant la recherche de documents sur la toile, selon des critères de recherche thématique et de difficulté. Or, il est évident que ce genre d'outils requiert l'insertion d'une formule de lisibilité en leur sein afin de remplir les tâches qui leur sont assignées.

Par ailleurs, une formule de lisibilité automatisée peut également être insérée au sein de logiciels d'ALAO, tels que des générateurs automatiques d'exercices. En effet, s'il existe quelques réalisations dans ce domaine pour le français (Chanier et Selva, 2000; Selva, 2002; Verlinde *et al.*, 2003), l'une de leurs limites actuelles est l'impossibilité de contrôler dynamiquement la difficulté contextuelle des matériaux utilisés comme base des exercices. À nouveau, l'ajout d'une formule de lisibilité au sein de ce type de logiciels pourrait rendre ceux-ci plus adaptables.

On voit donc qu'aussi bien les professeurs de langue que les concepteurs de systèmes d'ALAO pourraient trouver un intérêt à l'existence d'un modèle de lisibilité spécifique au FLE qui soit mature. Pourtant, très peu d'études existent sur la question et aucune d'elles n'intègre des technologies de TAL, d'où notre volonté d'investiguer cette problématique.

1. Pour être exact, il faut tout de même faire référence à Collins-Thompson et Callan (2005), qui ont adapté leur formule de lisibilité pour l'anglais L1 au français L1. Il s'agit néanmoins plutôt d'une étude de portabilité de leur modèle que d'une étude spécifique au français. Pour plus de détails, se référer au Chapitre 2.2.8.

Nous avons adopté pour ce faire une double démarche : théorique et critique, mais aussi exploratoire et expérimentale. Plus précisément, cette thèse s'attaque à un double objectif :

– déterminer si l'emploi de technologies de TAL dans le contexte de la lisibilité conduit bien une amélioration des performances par rapport aux formules plus classiques, utilisées dans les milieux de l'éducation ou du journalisme ;
– concrétiser ces apports, pour autant qu'ils soient avérés, au sein d'une nouvelle formule de lisibilité pour le FLE.

Ce programme est réalisé en trois grandes parties. La première, théorique, présente un ensemble de concepts issus des études sur la lecture nécessaires à la suite de l'étude, ainsi que les recherches précédentes en lisibilité. La seconde, méthodologique, met en place les outils et les données nécessaires à l'investigation de la thèse énoncée ci-dessus ; enfin, la dernière, expérimentale, rapporte les tests réalisés en vue de déterminer l'apport du TAL à la lisibilité, ainsi que la formule qui découle de ces expérimentations.

Plus précisément, la thèse s'organise selon cinq chapitres. Le premier chapitre aborde les processus de lecture en langues première et seconde selon un point de vue psycholinguistique. Si cette problématique ne ressortit pas directement à la lisibilité, il importe d'en présenter les grands concepts et les grandes tendances dont les répercussions sur les études en lisibilité sont majeures. C'est pourquoi ce chapitre débute avec une définition du processus de lecture tel qu'il est conçu actuellement, avant que ne soient envisagés un ensemble de dimensions de ce processus et une série de modèles théoriques ayant contribué à articuler ces dimensions et à les expliquer. Cette présentation se termine avec un rapide parcours de la façon dont la lecture a été envisagée au sein de l'enseignement du FLE, qui confirmera l'importance actuelle de cette compétence.

Dans le second chapitre, nous retraçons les grandes lignes du domaine de la lisibilité jusqu'à nos jours. Sachant que la dernière synthèse récente en français sur la question remonte à Henry (1975), il nous a semblé nécessaire de compléter ce travail en recensant les travaux plus récents, en particulier ceux qui appartiennent à ce nouveau mouvement qui allie le TAL et la lisibilité et que nous avons désigné sous le terme de « lisibilité computationnelle ». Bien qu'intéressé au premier plan par les études sur le français, notre propos met celles-ci en perspective par rapport à la masse des travaux sur l'anglais, également largement décrits. Nous espérons également, par ce parcours détaillé du domaine, raviver l'intérêt de la communauté scientifique pour la lisibilité du français.

Avec le troisième chapitre s'ouvre la partie méthodologique de la thèse. Celui-ci vise à établir, dans un premier temps, une nomenclature aussi exhaustive que possible des différentes variables qui ont été utilisées pour prédire la difficulté de textes en anglais ou en français.

Cette typologie est organisée en cinq grandes familles de prédicteurs, qui ressortissent aux dimensions lexicale, syntaxique, conceptuelle, organisationnelle des textes, ainsi qu'aux aspects spécifiques du processus de lecture en L2. Pour chacune des variables répertoriées au sein de ces familles, non seulement nous décrivons les différentes formes de paramétrisation dont elles ont fait l'objet précédemment, mais cherchons à rassembler une série d'arguments théoriques et empiriques attestant ou non de la réalité de leur effet sur le processus de lecture. Cet inventaire critique, au-delà de son intérêt intrinsèque, vise également à faire le tri parmi cet ensemble de prédicteurs, afin d'en retenir les plus pertinents. Ceux-ci sont listés dans la seconde partie de ce chapitre, lequel détaille aussi leurs procédures d'implémentation.

Quant au quatrième chapitre, il traite du processus d'acquisition de données annotées, lesquelles sont requises pour une approche de la lisibilité qui mêle des techniques de TAL à des algorithmes d'apprentissage automatisé issus de l'intelligence artificielle. Il propose tout d'abord une comparaison des différents critères qui peuvent être utilisés pour attribuer un niveau de difficulté de référence à un ensemble de textes et défend l'emploi de l'un d'eux en particulier : l'annotation par experts. Ensuite, sur la base de ce critère, il décrit la procédure de constitution du corpus de textes, annotés en termes de difficulté. Celui-ci fait enfin l'objet d'une analyse critique par rapport à certaines de ses caractéristiques susceptibles d'influencer les résultats de notre étude et qui découlent du critère retenu.

Enfin, le cinquième chapitre clôt cette thèse par la présentation des résultats des différents tests effectués en vue d'évaluer les apports du TAL à la lisibilité. Il rapporte dans un premier temps les analyses visant à déterminer la capacité de chaque prédicteur considéré seul dans sa relation à la difficulté des textes. Ces analyses permettent de mettre en évidence une série d'indices de la lisibilité particulièrement efficaces, mais également de prendre position par rapport à plusieurs problématiques transversales identifiées au cours de l'état de l'art. Ce chapitre s'achève sur des expérimentations aboutissent à la création d'une nouvelle formule de lisibilité pour le FLE, la première à intégrer des technologies de TAL pour classer des textes selon l'échelle du Cadre européen commun de référence pour les langues (CECR), une norme aujourd'hui généralisée dans l'espace européen.

Concluons cette introduction en soulignant que bien que la problématique de cette thèse s'insère dans un domaine particulier, la linguistique appliquée, et plus particulièrement, le TAL, nous avons sans cesse cherché à sortir du cadre de ces disciplines et à déborder sur d'autres domaines connexes ou plus éloignés. En effet, notre problématique constitue un phénomène si complexe qu'une seule des vues qu'offrent les sciences humaines n'aurait pas suffi à l'éclairer. C'est pourquoi nous avons récolté des faits au sein de la psychologie cognitive, de la didactique, de la neuropsychologie, de la sociologie et bien évidemment de la linguistique afin de les organiser en une vision que nous espérons aussi cohérente

que possible et plus informative qu'une approche purement disciplinaire. Sans compter que nous avons également employé nombre d'outils et techniques issus de la mathématique, de la statistique et de l'informatique. Cette multidisciplinarité résulte autant de notre conception personnelle de la recherche scientifique que d'une exigence issue de la problématique elle-même. En effet, comme le Chapitre 1 va le développer, la lecture est un phénomène complexe qui mobilise un grand nombre de chercheurs afin d'en éclaircir les mystères.

Première partie

Bases théoriques : lecture et lisibilité

CHAPITRE 1

LA LECTURE EN FLE

La semaine dernière, j'ai suivi un cours de lecture rapide — et ça a marché ! Hier, j'ai lu tout Guerre et Paix *en une heure ; C'est une histoire de Russes... (Woody Allen)*

1.1 Introduction

Dans l'introduction, nous avons montré que le sujet de cette étude ressortit au domaine de la lisibilité des textes. Sans trop anticiper sur le Chapitre 2 qui décrit cette notion en détail, on peut néanmoins déjà mettre en avant le fait que celle-ci s'appuie fortement sur la manière dont on perçoit la lecture. En effet, la lisibilité vise à associer au mieux un lecteur à un texte qu'il sera capable de lire, d'où la nécessité de mobiliser une série de concepts liés au déchiffrage et à la compréhension des textes. Autrement, comment évaluer la difficulté d'un texte si l'on ne sait pas d'où celle-ci provient ? Comment mesurer la compréhension d'un lecteur pour un texte donné, si cette notion de compréhension reste vague ?

Dès lors, il nous a semblé impossible de traiter de lisibilité sans articuler cette réflexion avec les connaissances actuelles sur les processus de la lecture en L1 et L2. En effet, depuis ses débuts, la lisibilité s'est nourrie de ces travaux à de nombreuses reprises pour enrichir sa propre réflexion. Toutefois, depuis les dernières réalisations importantes sur la lisibilité du français, la connaissance des processus de lecture a fait de nombreux progrès, avec l'arrivée de modèles computationnels simulant les processus cérébraux, mais aussi de diverses techniques d'imagerie mentale qui permet enfin une observation directe du lecteur en train de lire. C'est pourquoi le premier chapitre de cette thèse a été dédié à une synthèse des connaissances sur les processus de lecture en L1 et L2.

Cependant, compte tenu de la masse des recherches sur la lecture en langue première et en

langue seconde, il s'est avéré impossible de proposer ici un état de l'art complet sur la question [1]. D'ailleurs, certains auteurs jugent tout simplement cette tâche impossible. Alderson (2000, 1) affirme : « the sheer volume of research on the topic belies any individual's ability to process ». Cette assertion est confirmée par une étude de Urquhart et Weir (1998, 19) portant sur la période 1970-1995, qui estime à plus de 2500 le nombre d'articles comportant le terme « reading » dans leur titre publié annuellement.

Pour cette raison, nous avons dû opérer un choix parmi les modèles et théories existants. Notre objectif est simplement de proposer un résumé des connaissances actuelles, sur lesquelles s'appuient nos réflexions concernant la prédiction automatique de la difficulté des textes en FLE. Dès lors, cette synthèse débute par une définition de cette pratique complexe qu'est la lecture, articulée autour de trois pôles : le texte, le lecteur et la situation de communication. Ensuite, le processus de lecture est détaillé via la description de ses composantes principales : la perception visuelle, la reconnaissance des mots et la compréhension. Cette partie de l'exposé ne propose cependant pas de modèle théorique qui organise ces connaissances au sein d'une construction intellectuelle cohérente. C'est dans la troisième partie de ce chapitre que sont recensés les modèles majeurs qui ont postulé diverses organisations des processus de lecture. Soulignons par ailleurs que, dans ces trois sections, les faits rapportés concernent d'abord la lecture en langue maternelle, mais les particularités de la lecture en langue étrangère sont également mises en évidence.

Enfin, une fois clarifiés les mécanismes de lecture, nous faisons le pont entre la théorie et la pratique didactique et envisageons la place accordée à la lecture et aux textes écrits au sein de la didactique des langues étrangères. En effet, les études en lisibilité reposent sur le postulat que le texte occupe une place centrale dans l'enseignement des langues étrangères. Or, ce prérequis n'a pas toujours été rempli, selon la théorie de l'apprentissage ayant dominé les pratiques didactiques à une époque donnée. Nous présentons ainsi, dans ce dernier point, les différentes positions qui ont coexisté concernant la place de la lecture au sein de la classe de FLE et montrons les liens existant entre ces constructions pédagogiques et les postulats théoriques sur la lecture envisagés dans les premières parties de ce chapitre.

1. Le lecteur curieux peut tout de même consulter Ferrand (2007) pour une synthèse relativement complète des travaux récents sur les mécanismes de la lecture en langue première.

1.2 Les processus de lecture en langue première et seconde

> *Le spectacle de l'homme isolé dans sa lecture est un des plus trompeurs qui soient. De ce tête à tête avec le texte, il semble qu'il n'y a rien à dire, que toute étude, tout commentaire ne pourra jamais rendre compte du caractère singulier de cette alchimie où, par le moyen de quelque mystérieuse transmutation, des signes noirs inscrits sur le papier deviennent soudain images, rêves, pensées, actions et qu'il y aurait même quelque indélicatesse à vouloir pénétrer dans l'espace intérieur ainsi creé. (Vigner, 1979, 3)*

Les premiers travaux modernes sur la lecture en langues première et seconde remontent à la fin du 19e siècle. On compte parmi eux Javal et son *Essai sur la physiologie de la lecture* (1879), Cattell et son *The inertia of the eye and brain* (1885) et surtout Huey avec *The psychology and pedagogy of reading* (1908). À la suite des travaux de ces précurseurs, l'opinion se répand que les processus de lecture en L1 et en L2 sont similaires, c'est-à-dire, d'après Bernhardt (2003, 112) que, dans le cas de lecteurs expérimentés, on ne note guère de différence *en surface*. Cette dernière précision a toutefois tendance à être oubliée et l'on en vient à considérer que les processus internes doivent aussi être similaires.

Cette conviction a conduit à une forte dépendance des recherches sur la lecture en L2 vis-à-vis des travaux en L1, à tel point que Bernhardt (2005, 133) a pu dire : « much of the research of the 1980s could be in part characterized as the slavish replication of studies conducted in first language ». On trouve une belle illustration de cette tendance chez Alderson et Urquhart (1984, xv) qui, dans l'introduction de leur ouvrage intitulé *Reading in a Foreign Language*, défendent la position suivante :

> *Much of the research reported [here] (...) is from studies of first-language readers, which we consider to be of particular relevance to the foreign language reader. We do not, and indeed find it difficult to, draw a clear distinction between first and foreign language reading — in fact, it is not clear to what extend reading in a foreign language is different from reading in a first language.*

Cependant, peu à peu, les chercheurs se sont accordés à considérer la lecture en L2 comme une activité cognitive particulière. Bernhardt (2003, 112) explique : « the mere existence of a first-language renders the second-language reading process considerably different from the first-language reading process because of the nature of information stored in memory ». Pour Koda (2005, 4) également, les études sur la lecture en L1 constituent un point de départ

logique, mais ne sont pas capables de capturer les attributs particuliers de la lecture en L2.

C'est pourquoi, dans ce chapitre, nous présentons certes les conceptions développées à partir d'expérimentations sur la L1, mais en veillant aussi à mettre en évidence les particularités de la lecture en L2. Cette section s'ouvre avec un essai de définition de la lecture, qui présente les différentes facettes de cette activité : le lecteur, bien sûr, mais aussi le texte et la situation de lecture. Elle se poursuit par un examen des principales découvertes concernant les processus de lecture, organisés selon une approche simple que Urquhart et Weir (1998) désignent sous le terme de « modèles componentiels », c'est-à-dire une description indépendante des différentes composantes participant au processus de lecture. Cette approche est suivie d'un autre type de modélisation, les « modèles du processus de lecture », qui tentent d'organiser séquentiellement ou interactivement ces différentes composantes au sein d'une conception globale. C'est d'ailleurs par ce recensement des modèles dominants que nous refermons cette section.

1.2.1 Essai de définition : la lecture

Qu'est-ce que la lecture ? Difficile de répondre simplement à cette question sans déjà invoquer l'un ou l'autre concept, défendu par tel auteur et décrié par un autre. D'autant que ce domaine complexe est abordé selon différentes facettes par des chercheurs de tous horizons : des psychologues cognitifs, qui, depuis les années 60, ont abandonné le behaviorisme pour s'intéresser à nouveau aux processus de lecture ; des neurologues, qui visent à établir des hypothèses sur la base de cas pathologiques ou d'observations par imagerie cérébrale ; des didacticiens, qui cherchent plus à définir de bonnes pratiques qu'à théoriser le processus, des linguistes et même des ingénieurs linguistes. De plus, ainsi que Carver (1997, 4-5) l'a bien clarifié à travers sa formule, « one second, one minute and one year of reading », les chercheurs envisagent la lecture selon des temporalités diverses : la seconde durant laquelle l'oeil perçoit et traite un mot ou phrase ; la minute qui suffit à lire un court texte dont on peut ensuite mesurer le niveau de compréhension et l'année, qui relève de la question du développement des capacités de lecture à travers le temps. Dans cette thèse qui vise à mesurer la difficulté contextuelle de textes plutôt courts, nous nous limiterons à approcher la lecture selon les deux premières temporalités.

Pour commencer par une définition simple, reprenons la proposition de Grabe (2005, 51) : « reading can easily be defined simply as the ability to derive understanding from written text ». Il est toutefois légitime de se demander ce que cache l'expression « derive understanding ». Qu'est-ce au juste que la compréhension, et comment cette extraction du sens s'effectue-t-elle sur la base des graphèmes du texte ? Rayner et Pollatsek (1989, 23) précisent : « reading is the ability to extract visual information from the page and comprehend

the meaning of the text ». Dans cette définition apparait une étape préalable à l'interprétation du sens : l'extraction des données visuelles, qui correspond en gros à l'étape de reconnaissance des mots. La reconnaissance des mots et la compréhension forment ainsi le squelette de l'histoire de la recherche en lecture et on a pu observer, selon les époques, un mouvement de balancier dans l'importance respective accordée à ces deux composantes. Toutefois, bien d'autres aspects de la lecture ont également fait l'objet d'études. Urquhart et Weir (1998, 18) proposent une sorte de programme minimal pour tout bon modèle du processus de lecture :

> *We consider that any valid account of the reading process must consider such cognitive aspects as reading strategies, inferencing, memory, relating text to background knowledge, as well as decoding, and obvious language aspects as syntax and lexical knowlegde*.

Nous avons vu précédemment que ces différentes composantes listées par Urquhart and Weir pouvaient être décrites, soit séparément dans un modèle componentiel, soit de manière intégrée au sein d'un modèle du processus de lecture. Par ailleurs, une seconde dichotomie se dégage des travaux sur la lecture, à savoir que cette dernière est tantôt vue comme une habilité, tantôt comme une activité. Ainsi, les définitions de Grabe (2005) et Rayner et Pollatsek (1989) apparaissent représentatives de ce courant qui dépeint la lecture comme une habilité, même si d'autres auteurs préfèrent le terme de « compétence » (Koda, 2005). Cette conception largement répandue — sans doute sous l'action d'impératifs pédagogiques — présente l'avantage de justifier le décorticage de cette habilité en un ensemble de sous-habilités qui peuvent être explicitées et enseignées. Par contre, elle va souvent de pair avec la vision de ce que doit être un bon lecteur, voire avec l'idée qu'il existe un sens original du texte pouvant être extrait par le lecteur avec plus ou moins de succès [2]. Enfin, si le sujet qui manie cette compétence est le centre de tous les intérêts et que le texte écrit, présent dans les définitions ci-dessus, est également considéré, la dimension situationnelle de la lecture — et donc sa fonction de communication — nous semble occultée.

C'est pourquoi, Giasson (1997) [3], Widdowson (1979) [4] ou Rayner *et al.* (2001) [5], considèrent plutôt la lecture comme un processus, c'est-à-dire la suite des différentes phases d'un phénomène, ou une activité. Dès lors, elle met en jeu un **acteur**, le lecteur, et un **objet**, le texte écrit, dans un **contexte** donné (Deschênes, 1988; Giasson, 1990), trois facteurs qui méritent d'être précisés.

2. C'est le reproche que font Urquhart et Weir (1998, 87) à Rayner et Pollatsek (1989).
3. La lecture est pour Giasson « un **processus** dynamique » (Giasson, 1997, 21).
4. Widdowson (1979) cité par Alderson et Urquhart (1984, xxv) définit ainsi la lecture : « the **process** of getting linguistic information via print ».
5. Ces auteurs envisagent plusieurs définitions, et s'arrêtent finalement sur la suivante : « the **process** of gaining meaning from print » (Rayner *et al.*, 2001, 34).

Le lecteur

Le lecteur constitue la variable principale dans l'activité de lecture. En effet, il est commun de considérer qu'il y a autant d'interprétations d'un texte qu'il y a de lecteurs, puisque chacun d'eux va l'aborder en fonction de ses propres particularités. Pour Giasson (1997), ces particularités se situent à deux niveaux : les processus et les structures. Nous ne traiterons pas ici des processus, puisqu'il s'agit de la thématique centrale de la prochaine section. Quant aux structures, Giasson distingue les structures cognitives et affectives. Les premières comprennent « les connaissances que le lecteur possède sur la langue [à savoir les connaissances phonologiques, syntaxiques, pragmatiques, sémantiques] et sur le monde » (Giasson, 1997, 33). Les secondes déterminent la manière dont le lecteur aborde la tâche de lecture et qui est fonction de son attitude générale vis-à-vis de la lecture, mais aussi de sa motivation du moment, de sa familiarité avec le sujet du texte ou de l'intérêt qu'il lui porte. Toutefois, certains chercheurs actuels (Dehaene, 2007; Rayner *et al.*, 2001) semblent estimer que les structures, si elles influencent clairement le résultat de la lecture, relève probablement de tâches cognitives globales telle la compréhension du langage ou la résolution de problème.

Ces diverses particularités définissent certaines catégories de lecteurs : enfants, adolescents ou adultes ; issus d'un milieu socio-économique favorisé ou dévaforisé ; lettrés ou illettrés ; d'intelligence normale, aiguë ou au contraire atteints de retards mentaux... Alors que les pédagogues de la lecture ont tendance à insister sur les effets de ces spécificités, les psychologues de la lecture s'attachent plutôt à dépeindre chez le lecteur adulte des mécanismes universels, relativement similaires malgré les particularités de chacun. Ainsi, Rayner et Pollatsek (1989, 460) estiment qu'il est raisonnable de spécifier un modèle général unique pour l'ensemble des lecteurs compétents

Le lecteur qui nous intéresse ici est toutefois différent de ce lecteur adulte qui opère dans sa langue maternelle. En effet, le lecteur moyen dans une L2 se démarque de son homologue en L1 par les trois faits suivants (Koda, 2005, 7) :

1. Le plus évident est l'existence, en arrière fond, d'une autre langue qui a déjà profondément marqué les systèmes cérébraux de l'apprenant et a engendré des automatismes susceptibles de faciliter ou de contrarier sa lecture en L2. On pense ici aux congénères, aux phénomènes d'interférences, etc.
2. Souvent plus matures que des jeunes enfants apprenant à lire en L1, les lecteurs en L2 peuvent prendre appui sur leurs expériences préalables, en matière d'alphabétisation, de réseau conceptuel et de connaissances du monde. Ainsi par exemple, l'apprentissage du vocabulaire se limite bien souvent à l'association d'une seconde étiquette à un concept déjà bien développé.
3. Par contre, à la différence du lecteur en L1 qui se repose sur ses connaissances linguis-

> tiques acquises à l'oral [6], l'apprenant en L2 ne peut se reposer sur la même masse de connaissances langagières, l'apprentissage de la lecture en L2 commençant souvent peu après — quand il n'est pas concomitant avec — la découverte de la L2. Ajoutons à ce déficit linguistique, un possible déficit touchant les connaissances culturelles partagées au sein de la culture associée à cette L2.

La différence principale entre la recherche en L1 et en L2 est ainsi résumée par Bernhardt (2003, 113) : « [L2 research] must be able to explain cognitive and social literacy processes in light of a literacy/language network that already exist ». Or, selon la culture et la langue d'origine des apprenants, ce réseau est susceptible de fluctuer. Peut-on dès lors, semblablement à ce qui se fait pour le lecteur adulte en L1, postuler une simulitude des mécanismes qui justifie un modèle unique, ou faut-il modéliser ce groupe de lecteurs en tenant compte de son hétérogénéité, comme le préconise Bernhardt (1991; 2003; 2005) ?

À l'origine, les processus de lecture ont été jugés universels quelle que soit le background des apprenants, puis, entre le milieu des années 70 et celui des années 90, une série de linguistes appliqués (Alderson et Urquhart, 1984; Bernhardt, 1991; Carrell *et al.*, 1989; Mackay *et al.*, 1979) ont mis en évidence l'importance de la langue maternelle des apprenants comme source d'hétérogénéité, mais le lecteur modélisé restait un adulte sachant déjà lire dans sa langue. À la fin des années 90, Bernhardt et Kamil (2006, 88) rapportent que la présence accrue d'immigrants de seconde génération dans le système scolaire a conduit les chercheurs à prendre conscience d'une autre source d'hétérogénéité : le degré d'alphabétisation de l'apprenant dans sa propre langue.

Cette conception complexifie cependant considérablement la modélisation des processus de la lecture, ce qui explique peut être pourquoi on observe une carence à ce sujet dans la littérature. Pour notre propos, nous nous limiterons à présenter des faits concernants l'apprenant de FLE adulte ou grand adolescent moyen, en proposant à l'occasion des perspectives issues de l'approche *cross-linguistique* représentée par (Bernhardt, 2005; Koda, 2005; Steffensen *et al.*, 1979).

Le texte

Avant tout, il convient de préciser ce que nous entendons sous le terme *texte*. Pour Urquhart et Weir (1998, 14), la lecture s'applique plus exactement à des messages langagiers sous une forme écrite ou imprimée. Mais, que cache le terme *messages langagiers* ? Autrement

6. Grabe (1991, 386) rapporte : « First language learners have already learned somewhere on the order of 5,000 to 7,000 words before they formally begin reading instruction in schools (Singer, 1981). They also have a good intuitive sense of the grammar of the language. L2 learners typically have not already learned a large store of oral language vocabulary ; nor do they have a fairly complete sense of the grammar of the language ».

dit, peut-on *lire* du code informatique, une partition ou une formule mathématique ? Bien que Urquhart et Weir (1998, 15) rejettent ces types de langages formels hors du champ de la lecture, il nous apparaît qu'il s'agit bien d'un processus correspondant à une extraction de sens depuis une forme écrite de langage. Cependant, de récentes découvertes en neuropsychologie [7] laissent à penser que les réseaux neuronaux mis en œuvre pour la lecture de langages formels diffèrent de ceux qui s'appliquent à la lecture des langages naturels. Dès lors, dans la suite de cette thèse, nous envisagerons la lecture comme une activité qui s'applique uniquement à des textes écrits rédigés en langage naturel.

Cette question clarifiée, il reste à définir les caractéristiques du texte qui influencent la lecture. Elles occupent une place privilégiée au sein de nos préoccupations, puisque la lisibilité s'intéresse avant tout à la variable texte, cherchant à en quantifier la difficulté pour un public déterminé. Deschênes (1988) classe les caractéristiques textuelles en deux grandes catégories, reprenant la distinction traditionnelle entre la forme et le contenu. Par contenu, il entend l'aspect sémantique du texte, c'est-à-dire, son contenu informationnel qui peut être décrit à l'aide de deux notions : la microstructure, à savoir la structure des propositions constitutives du texte et leurs relations (Kintsch et van Dijk, 1978, 365), et la macrostructure, qui est une structure de signification globale construite sur la base des propositions du texte. Sous les aspects formels, il envisage la présence de mots charnières dans le texte et le niveau de redondance syntaxique, mais — sans doute influencé par l'importance des recherches sur la structure textuelle à cette époque — il ne considère pas des choses aussi essentielles que le vocabulaire ou les structures lexicales employés.

Nous traitons ces différentes variables textuelles plus en détail dans le chapitre 3. Sur la base d'études empiriques issues de la littérature et de nos propres expérimentations, nous verrons l'importance que ces variables peuvent avoir sur la difficulté que présente un texte donné pour l'apprenant qui le lit. Toutefois, étant donné qu'il s'agit là d'un aspect central de ce travail, nous ne l'anticiperons pas davantage dans cette section.

La situation de communication

Dans son fameux *Understanding reading*, Smith (1971, 12) écrit : « Reading is an act of communication in which information is transferred from a transmitter to a receiver », adaptant à la lecture le modèle de la communication proposé par Shannon et Weaver (1949). Cette conceptualisation de la lecture a l'avantage de formaliser au sein d'un paradigme uni-

7. Dehaene (2007, 91) rapporte le cas de patients atteints d'alexie, une maladie qui touche les centres de la reconnaissance des mots, et précise que ceux-ci sont toujours capables de traiter les chiffres et les nombres. Il en conclut une séparation claire quant aux zones du cerveau chargées de traiter ces deux types de langages.

fié plusieurs faits [8], ainsi que d'insister sur le rôle actif du lecteur, qui doit décoder le message qu'il reçoit. Le modèle de Shannon, appliqué à la lecture, présente néanmoins certains défauts, dont le principal est sans doute qu'il ne distingue pas explicitement les rapports scripteur/lecteur, scripteur/texte et texte/lecteur (ce que fait par contre Moirand (1979, 10) dans son propre schéma de la situation d'écrit), mais également qu'il propose une vision de la communication à un seul canal, ne permettant pas d'éclairer correctement la situation de communication propre à cette relation particulière entre le lecteur et l'auteur.

Cette relation semble plus justement caractérisée par Vigner (1979, 13) comme une « situation de communication différée », où un laps de temps plus ou moins long sépare la production du message de sa réception. Cela signifie que, dans la plupart des cas, l'auteur est absent quand le lecteur découvre le texte. Dès lors, à la différence d'une interaction verbale, il ne peut adapter son discours (au niveau de la difficulté linguistique, du contenu, etc.) en fonction des réactions du lecteur : il lui faut imaginer la manière dont son texte va être compris et « pourvoir son message d'un système d'autocontrôle de la réception » (Vigner, 1979, 12). De même, le lecteur ne peut ni poser des questions pour éclaircir un point du discours qui lui reste obscur, ni tirer profil des gestes, expressions faciales, intonations, rythme, bref de cet orchestre — ainsi que l'on appelé par analogie les adeptes de la nouvelle communication (Winkin, 1981) — que lui joue son interlocuteur afin de faire passer au mieux le message [9]. À la place, il doit se contenter de divers éléments de mise en pages, tels les titres et sous-titres, des marques de mise en évidence, ou encore des images, des tableaux ou des graphiques, qui visent à renforcer le message contenu dans le texte écrit.

Une seconde caractéristique de cette communication différée est que le texte ne se limite généralement pas à un lecteur unique, mais cible un lectorat plus ou moins étendu et plus ou moins connu du scripteur. Ajoutons que les interlocuteurs ne partagent pas la même situation d'énonciation et que « le texte se présente au lecteur hors de sa situation d'origine » (Jouve, 1993, 13). L'auteur gagne dès lors à expliciter au maximum le cadre de référence de son message afin que les déictiques et autres marques situées puissent s'interpréter en fonction de la situation d'écriture et non de la situation de lecture. Enfin, cet étalement de la communication à travers le temps offre l'avantage, aussi bien au rédacteur qu'au lecteur, de réguler

8. Pour plus de détails sur la lecture vue sous l'éclairage de la théorie de l'information, consulter Smith (1971, 12-27). Parmi les perspectives apportées par Smith, citons les limites de la perception visuelle analysées sous l'angle de la capacité du canal de communication ou encore la sélection d'alternatives opérées, tant au niveau de l'identification des lettres ou des mots qu'au niveau du sens, grâce à l'apport d'information (Goodman (1967, 1968) parle pour sa part de « guess » sur la base d'indices).

9. Vigner (1979, 12) décrit ainsi tout le drame de la lecture : « Le malentendu est donc au coeur de la communication. Produire un texte écrit, ce sera formuler une hypothèse sur les conditions de réception et d'appréciation du message qu'on souhaite transmettre. Lire un texte écrit consistera à s'interroger sur le sens exact du message que l'on reçoit, sur les intentions de son auteur. La communication écrite ne fera donc que mettre en relation deux incertitudes, celle du scripteur quant aux réactions et au comportement de son lecteur, et celle du lecteur relative aux intentions du scripteur. »

le débit des informations. L'auteur, qui ne court pas le risque d'être interrompu par son interlocuteur, peut prendre le temps de construire son discours : ses phrases sont normalement plus structurées et plus complexes, d'où le fait que l'écrit se révèle être un cadre propice à l'emploi de certaines variations linguistiques peu rencontrées à l'oral [10]. Le lecteur, quant à lui, peut s'octroyer le luxe de s'arrêter pour réfléchir à ce qu'il a lu, retourner en arrière pour se remémorer quelque chose ou prendre des notes.

Au niveau de la relation entre le lecteur et le texte, la vision de Smith (1971) qui présente la lecture comme une activité linéaire où le lecteur est actif ne fait plus l'unanimité. Les chercheurs se figurent plutôt ce rapport comme interactif (Carrell *et al.*, 1989; Kintsch et van Dijk, 1978; Stanovich, 1980), car le lecteur ne se contente pas d'éclairer le contenu du texte à la lumière de ses connaissances préalables, mais s'engage dans un véritable processus interactif de construction du sens où se mêlent les informations en provenance du texte et ses propres connaissances. Dans cette perspective, le modèle cybernétique de la communication de Wiener (1948, cité par Winkin, 1981) semble convenir davantage que le modèle de Shannon pour décrire ce rapport du lecteur au texte. En effet, le concept de feedback qu'il postule rencontre celui d'interaction en dépeignant un lecteur capable d'analyser le résultat de sa lecture et d'adapter en conséquence ses stratégies afin d'atteindre ses objectifs.

À ce propos, l'objectif avec lequel un texte est abordé a aussi été considéré comme un facteur contextuel influençant d'une part les informations extraites du texte [11], de même que le type de lecture utilisée par le lecteur. Malheureusement, les auteurs ne s'accordent pas complètement sur les différents types de lecture qui existent, ni sur leurs définitions. L'un des principaux, Carver (1997, 5-7) distingue :

1. le **balayage** (« scanning ») qui consiste à parcourir le texte de manière sélective pour y trouver une information spécifique ou s'en faire une impression initiale ;
2. l'**écrémage** (« skimming ») qui revient à survoler le texte pour en extraire les idées principales ;
3. la **lecture ordinaire** (« rauding ») ;
4. **l'apprentissage** qui demande d'intégrer un ensemble d'informations issues du texte en mémoire à long terme ;
5. la **mémorisation**, assez similaire à l'apprentissage, mais auquel s'ajoute l'idée d'une répétition des faits contenus dans le texte afin de les mémoriser.

10. Citons comme exemples pour le français des formes lexicales telles que *nonobstant*, *il appert* ou des temps verbaux tels que le passé simple ou le subjonctif imparfait

11. Anderson et Pichert (1978) ont donné à lire à deux groupes le même texte décrivant une maison assez vieille, mais richement meublée. Les membres du premier groupe devaient se mettre dans la peau d'un cambrioleur et ceux du second s'intéresser à la maison du point de vue d'un acheteur potentiel. L'expérience a prouvé que, lorsque les lecteurs doivent se souvenir du contenu d'un texte, on observe que la perspective de lecture influence aussi bien l'intégration du contenu que sa récupération en mémoire.

Il organise ces formes de lecture de manière décroissante sur un continuum, en fonction de la quantité de texte qu'elles permettent de traiter en moyenne (en nombre de mots par minute). Ainsi, plus la tâche est complexe et plus la vitesse de lecture est réduite.

Cette question du type de lecture pratiquée est d'une importance particulière dans le contexte L2, où d'autres formes que la lecture ordinaire sont envisagées. En effet, des techniques telles que l'écrémage ou le balayage sont de plus en plus utilisées dans nos sociétés modernes (Urquhart et Weir, 1998, 101). Dès lors, comme l'explique Nuttall (1996, 49), il importe de stimuler chez les apprenants une certaine flexibilité dans leurs pratiques, afin que ceux-ci soient capables d'adapter leurs stratégies de lecture en fonction des objectifs concrets qu'ils poursuivent. C'est pourquoi les stratégies de balayage et, surtout, d'écrémage sont mises en pratique au sein des méthodes d'enseignement du FLE actuelles (*cf.* Section 1.3), et en particulier dans la réalisation des tâches propres à la perspective actionnelle. Pourtant, c'est la lecture ordinaire qui fait traditionnellement — et presque exclusivement — l'objet des études en psychologie cognitive est la lecture ordinaire.

Conclusion

Cette courte réflexion sur la nature de la lecture a mis en évidence la difficulté d'arriver à une description définitive de cette activité complexe, tant les variables qu'il faut prendre en compte sont nombreuses et liées entre elles. Mais surtout, c'est leur importance relative qui est difficile à estimer, d'autant qu'elle varie à chaque lecture. En effet, pour un lecteur donné, ce seront avant tout les aspects lexicaux et syntaxiques qui nuiront à sa compréhension. Un autre, éloigné temporellement ou spatialement du texte, manquera de connaissances encyclopédiques pour suppléer les informations implicites nécessaires à la compréhension, tandis qu'un troisième, victime d'une dyslexie, peinera dans le décodage des mots ; qu'un quatrième, stressé par une situation d'examen, en viendra à pratiquer une lecture hésitante, etc.

Confronté à cette complexité, nous avons décidé de limiter notre propos à un certain type de lecteur, de texte et de situation de communication. Dans la suite de ce chapitre, nous chercherons à décrire un lecteur de FLE adulte ou grand adolescent, dont les capacités cognitives sont normales et qui pratique, dans un environnement d'apprentissage, généralement une classe, essentiellement une lecture ordinaire ou un écrémage de texte. Cette option de se limiter principalement à ces deux types de lecture s'explique par les objectifs de cette recherche. D'une part, comme Koda (2005, 6) l'a souligné, la mesure d'évaluation change en fonction de l'objectif : ce qui compte pour le repérage d'une information (balayage), c'est la vitesse, tandis pour l'extraction des idées principales d'un texte (écrémage) ou sa lecture attentive (*rauding*), c'est davantage la qualité de la compréhension qui importe. Or,

c'est justement cette seconde dimension qui est la plus souvent mesurée dans les études en lisibilité. D'autre part, il nous semble, en accord avec Grabe (2005, 50) que l'habilité la plus fondamentale pour la lecture en L2 est bien la compréhension du contenu principal d'un texte et que, dès lors, une formule de lisibilité doit avant tout viser la lecture normale ou, à la limite, l'écrémage, puisque celui-ci vise aussi, et uniquement, à dégager l'idée principale d'un document (Richards et Schmidt, 2002, 467).

Dans ces deux types de lecture, nous avons vu que la reconnaissance des mots et la compréhension en constituent des pierres angulaires. Pourtant, nombre d'autres opérations se déroulent également pendant la lecture, et ce d'autant plus lorsqu'elle s'applique à une langue étrangère. Les descriptions de ces mystérieuses opérations ont connu un développement important au cours des cinquante dernières années, passant tantôt par la défense d'une série de conclusions, tantôt de leurs inverses. Peu à peu, ce mouvement de balancier a permis de dégager un ensemble de faits concernant le processus de lecture en L1 et, dans une moindre mesure, en L2. La suite de cette section présentera une synthèse des problématiques essentielles de la lecture avant d'aborder les différents paradigmes théoriques qui ont dominé les recherches.

1.2.2 Les grandes problématiques de la lecture : une approche componentielle

Lorsqu'on cherche à dépeindre les connaissances actuelles concernant la lecture en L1 et en L2, on est confronté à un problème d'organisation. Entre la perception visuelle et l'interprétation des mots qui viennent d'être lus, ne s'écoulent que quelques dixièmes de secondes. Pourtant, durant cette courte période, de multiples opérations cognitives ont lieu. Ne suffit-il pas dès lors de les décrire les unes après les autres ? Certes, mais cette façon de faire comporte deux obstacles non négligeables. D'une part, les auteurs ne s'accordent pas sur un même ensemble de composantes, ni sur une limite claire entre celles qui sont spécifiques à l'activité de lecture et celles qui relèvent d'activités cognitives plus générales telles la perception visuelle ou la compréhension du langage. D'autre part, cette approche componentielle, en évitant de décrire la manière dont les diverses composantes échangent de l'information, reste largement sous-spécifiée, puisqu'il est aujourd'hui largement accepté que ces composantes interagissent.

L'approche componentielle peut donc être considérée comme un modèle sous-spécifié des processus de la lecture. Nous l'emploierons toutefois dans cette section, car elle permet d'introduire les problématiques essentielles de la lecture sans se préoccuper encore de la manière dont ces différents processus interagissent les uns avec les autres. Il s'agit d'ailleurs d'un type de présentation fréquemment utilisé par les chercheurs originaires du champ de

la linguistique appliquée [12] ou de l'acquisition des langues étrangères, parce qu'il présente l'avantage, non négligeable, de rester au plus près des faits empiriques et de ne pas postuler d'hypothèses fortes comme c'est le cas dans les modèles théoriques envisagés dans la section 1.2.3.

Comme exemple de modèle componentiel, on peut citer la « simple view » de Hoover et Gough (1990) et Hoover et Tunmer (1993). Ces auteurs ont formulé pour la lecture en L1 une conception simple ne comprenant que deux composantes : la reconnaissance des mots (le décodage) et la compréhension linguistique. Pour opérer une telle simplification, ils partent du principe qu'il convient de distinguer les habilités spécifiques à la lecture de celles qui interviennent dans la compréhension de l'écrit et de l'oral en général (Hoover et Gough, 1990, 128). Ils définissent ainsi le décodage (Hoover et Gough, 1990, 130) [13] :

> *[it is] the ability to rapidly derive a representation from printed input that allows access to the appropriate entry in the mental lexicon, and thus, the retrieval of semantic information at the word level.*

Quant à la compréhension linguistique, elle est définie ainsi (Hoover et Gough, 1990, 131) :

> *[it is] the ability to take lexical information (i.e., semantic information at the word level) and derive sentence and discourse interpretations.*

Notons également que les auteurs respectent une condition primordiale d'un modèle componentiel : la séparabilité des composantes. Ils en avancent comme preuve l'existence de dyslexiques, qui comprennent parfaitement le langage, mais éprouvent des difficultés au décodage, ou d'hyperlexiques, présentant les symptômes inverses [14].

Un autre exemple de modèle componentiel, cette fois dans le domaine de la lecture en L2, est celui de Coady (1979). Il comprend quant à lui trois composantes :

– les habilités conceptuelles (« Conceptual abilities ») : cette capacité d'abstraction, liée aux facultés intellectuelles, est importante pour parvenir à lire couramment, particulièrement dans le cas d'une langue étrangère. De plus, si elle pose problème, on peut difficilement y remédier (Coady, 1979, 7).
– les connaissances préalables (« Background knowledge ») : dans la lignée du modèle psy-

12. Sur les différences d'approche entre les chercheurs issus de la psychologie cognitive et ceux issus de la linguistique cognitive et sur les difficultés que ces derniers ont rencontrées dans leur tentative de constituer des modèles de la lecture, le lecteur curieux consultera Urquhart et Weir (1998, 234–241).

13. Hoover et Tunmer (1993) distinguent les termes *reconnaissance des mots* et *décodage* : le premier désigne l'accès général au lexique, tandis que le second représente l'accès par la voie phonologique, qui est donc la seule voie d'accès au lexique mental prise en compte dans le modèle. Les auteurs se justifient de cette restriction en arguant de l'emploi de leur modèle pour des lecteurs débutants. Nous ne reprenons pas cette distinction dans la suite de notre étude, considérant les deux termes comme des synonymes.

14. Pour d'autres faits empiriques attestant de cette séparabilité, le lecteur se reportera à Hoover et Gough (1990, 128-129). Pour une critique de ce modèle, il pourra consulter Urquhart et Weir (1998, 47-49).

cholinguistique de Goodman (1967) (*cf.* Section 1.2.3) dont s'inspire Coady, ces connaissances qu'a le lecteur sur le monde réel, mais aussi sur les règles de structuration de l'écrit, influencent considérablement la lecture, et ce d'autant plus que l'apprenant n'est pas issu de la même zone culturelle que le texte qu'il lit.

– les stratégies procédurales (« Process strategies ») : elles rassemblent l'ensemble des composantes de la compétence de lecture, parmi lesquelles Coady (1979, 7) cite les réseaux de correspondances entre les graphèmes et les phonèmes, ainsi qu'entre les syllabes et les morphèmes ; les informations sur la syntaxe, le sens des mots et les contextes ; les stratégies cognitives ou encore les motivations affectives.

On note donc que ce second modèle componentiel ne s'accorde pas du tout avec la segmentation de Hoover et Gough (1990). Il considère davantage de facteurs et les organise selon un schéma triangulaire, qui favorise nettement les composantes de niveau supérieur, au détriment de la reconnaissance des mots.

Ces deux exemples mettent ainsi en exergue la difficulté d'arrêter un choix vis-à-vis des composantes qui participent effectivement au processus de lecture et qui doivent être, dans une certaine mesure, indépendantes les unes des autres. Les récents progrès effectués en imagerie mentale (*cf.* Section 1.2.3) permettent cependant d'éclaircir ce point délicat. En effet, il est désormais possible d'identifier les aires du cerveau actives dans la lecture et de leur associer, avec plus ou moins de certitude, une fonction. D'après les résultats d'études en neuropsychologie cognitive (Dehaene, 2007; Salmelin et Kujala, 2006), il est possible de dresser un tableau des processus de lecture qui postule les trois composantes suivantes :

– la perception visuelle : s'il ne constitue pas un processus propre à la lecture, l'étude du mouvement des yeux se révèle très intéressant pour l'investigation des processus de lecture, ainsi que pour la problématique de la difficulté d'un texte.
– la reconnaissance des mots : il s'agit certainement de la composante la plus particulière à la lecture, comme l'ont défendu Hoover et Gough (1990), mais aussi Rayner *et al.* (2001) ou encore Plaut (2005).
– la compréhension du sens : elle dépend de circuits neuronaux plus généraux dédiés au traitement du langage oral et écrit (Smiley *et al.*, 1977), même s'il est fort probable qu'apparaissent des variations dues aux particularismes de la situation de communication propre à la lecture.

Nous avons choisi d'organiser la suite de notre propos, dans lequel sera présenté un ensemble de faits connus à propos des processus de lecture, en fonction de ce découpage en trois composantes.

La perception visuelle pendant la lecture

Première des étapes qui conduisent l'homme à appréhender une série de traits et de points pour en extraire de l'information, la perception visuelle ne relève pas des processus propres à la lecture, mais plutôt de l'activité perceptive en générale. Ce sont principalement les mouvements des yeux pendant la lecture qui ont fait l'objet de nombreuses études, et cela, depuis la fin du 19e siècle. Ainsi, Javal (1879, cité par Venezky 1984), aidé par Lamarre, serait le premier a avoir décrit le mouvement des yeux lors de la lecture dans *Physiologie et lecture de l'écriture*. Durant cette époque, qui s'étend jusqu'aux travaux de Huey (1908, cité Venezky 1984), plusieurs découvertes essentielles sont effectuées, alors que les techniques d'investigation sont très simples. Javal, par exemple, fait un trou dans un journal afin d'observer les yeux du lecteur pendant que celui-ci le parcourt.

Aujourd'hui, la technologie a bien évolué : il est possible de mesurer précisément les mouvements des yeux à l'aide des trois techniques d'oculométrie suivantes (Ferrand, 2007, 30; Rayner et Juhasz, 2006, 374) : une série d'électrodes posées près des yeux qui mesurent des variations de potentiels électriques de surface, l'analyse d'enregistrements vidéo de la pupille ou l'envoi d'un faisceau de lumière infrarouge au centre de la pupille afin d'en détecter le centre. Toutefois, la psychologie cognitive ne se limite pas à décrire ces mouvements per se, à la différence de l'approche behavioriste (Rayner, 1998, 372). Leur étude constitue également une voie d'accès aux processus internes qui ont lieu lors de la lecture (cf. Just et Carpenter, 1980; Rayner et Juhasz, 2006; Rayner *et al.*, 2005). Toutefois, dans cette section, nous nous limiterons à rapporter une série de faits concernant les mouvements visuels, sans envisager leur utilité méthodologique pour l'étude de la lecture.

Contrairement à ce qui a longtemps été admis, le mouvement oculaire n'est pas linéaire, mais opère par saccades : les yeux appréhendent une ligne de texte en un certain nombre de fixations, lequel est fonction de la compétence du lecteur. À l'époque de Javal, la controverse portait sur la question de savoir si la perception a lieu lorsque les yeux bougent ou lors des temps de fixation (Venezky, 1984, 9). Aujourd'hui, on sait que c'est lors des fixations que s'effectue la perception visuelle et on distingue habituellement trois types de « mouvements » oculaires :

1. Les mouvements de progression ou saccades : il s'agit de mouvements balistiques de l'oeil qui entraînent un changement du point de fixation. Ils s'effectuent très rapidement : leur durée est estimée à 25 millisecondes par Richaudeau (1969, 21), entre 10 et 23 ms par Gough (1972, 332) et entre 20 et 40 ms par Rayner *et al.* (2001, 46). Rayner (1998, 373) rapporte que leur vitesse dépend de la distance couverte entre les deux points de fixation et tourne autour de 30 ms pour la lecture, étant donné que la distance est très courte. En effet, la longueur moyenne d'une saccade correspond à un

angle oculaire de 2 ou 3˚ou à un déplacement de 7 à 9 lettres (Rayner *et al.*, 2005, 80; Dehaene, 2007, 38) , mais elle connaît une variance importante [15].

Ce besoin constant de déplacer l'oeil est dû à la structure de l'oeil. L'acuité visuelle n'est pas uniforme : elle est optimale au centre, en un lieu appelé la fovea, qui est riche en cônes (des cellules photoréceptrices hautement performantes), mais diminue dans la région parafovéale et plus encore dans la région périphérique où les récepteurs ne sont plus capables de discriminer correctement les lettres. Comme la taille de la fovea est réduite, l'oeil tend à bouger régulièrement afin de placer le texte à lire en face de celle-ci (Rayner, 1998, 374).

Durant ce mouvement très rapide, l'oeil est quasiment aveugle. On nomme ce phénomène la « suppression saccadique » et il s'expliquerait par la stimulation rétinale générée par le mouvement rapide de l'oeil (Matin, 1974, 904). Pour Campbell et Wurtz (1978), la rétine reçoit bien une image floue durant les saccades, mais celle-ci est masquée par la présence d'une image statique avant et après le mouvement. Par ailleurs, alors que les chercheurs tendent à considérer que les processus cognitifs sont interrompus durant les saccades (Rayner, 1998, 373), Irwin (1998) a montré que ce n'était pas le cas, au moins pour le traitement lexical durant la reconnaissance des mots.

2. Les fixations : elles correspondent à une courte période au cours de laquelle les yeux restent immobiles et la perception visuelle a lieu. Leur durée moyenne se situe entre 250 et 333 ms d'après Richaudeau (1969, 21), entre 200 et 300 ms d'après Rayner (1998, 373) [16], mais Just et Carpenter (1980, 330) ont montré que cette durée est sujette à une grande variance. Ils rapportent des fixations dépassant la seconde et demie et d'autres réduites à 83 ms [17]. Toutefois, quelques expériences laissent penser que ce temps n'est pas utilisé uniquement pour la perception visuelle, puisque 50 ms de présentation suffisent pour reconnaître un mot, sans que le traitement de l'information perçue durant cette période soit terminé après ces 50 ms (Dehaene, 2007, 41).

   Une autre variable importante est le nombre de lettres perçues lors d'une fixation, qu'on appelle l'empan perceptuel. Il est légèrement plus grand que la distance entre deux saccades (il y a donc un certain recouvrement entre deux empans consécutifs). Sa longueur a été estimée à 10 lettres par Javal (Venezky, 1984, 9), puis étendue à 17-18 caractères par Gough (1972, 332). C'est surtout grâce à la technique de la fenêtre mobile développée par McConkie et Rayner (1975) qu'ont été récoltées les données les plus intéressantes à ce sujet. Ainsi, McConkie et Rayner (1976), en recourant à

15. Se reporter aux distributions présentées par Rayner (1998, 376).

16. Plus exactement, Rayner (1998, 373) rapporte que le temps de fixation moyen diffère pour la lecture silencieuse (225 ms) et pour la lecture orale (275 ms). De même, la distance moyenne parcourue pendant chaque saccade est respectivement de 8 et 6 lettres.

17. La distribution présentée par Rayner (1998, 376) est toutefois plus homogène, puisque les valeurs oscillent plutôt entre 100 et 350 ms.

cette technique, ont mis en évidence l'asymétrie de l'empan perceptuel pour l'anglais, qui s'étend de 3 à 4 caractères à gauche et de 14 à 15 lettres à droite. Cette asymétrie est très clairement liée à la direction de la lecture, puisqu'elle s'inverse dans le cas de langues comme l'hébreu ou l'arabe où l'on lit de droite à gauche (Rayner, 1998, 380). On notera également que les particularités des langues influencent la taille de l'empan. Ainsi, en chinois ou en japonais, parce que la densité des caractères est supérieure à l'anglais ou le français, l'empan est réduit d'autant (cf. Rayner, 1998, 380).

Par ailleurs, il convient de distinguer l'empan perceptuel décrit ci-dessus et l'empan de reconnaissance des mots, dont les mots sont effectivement reconnus et traités par les processus sémantiques. Underwood et McConkie (1985) ont établi que ce dernier est plus réduit et n'excède généralement pas 7 à 8 caractères sur la droite. Les lettres situées juste au-delà sont distinguées de manière assez grossière par la région parafovéale et leur perception est supposée fournir des indices sur l'endroit où effectuer la prochaine saccade tout en initiant l'identification du mot suivant (Rayner, 1998, 381). Il arrive d'ailleurs que certains mots, complètement identifiés dans la région parafovéale, soient dès lors sautés lors de la saccade suivante. Ainsi, Just et Carpenter (1980, 330) indiquent qu'un lecteur normal confronté à un texte de difficulté adéquate lit en moyenne 1,2 mot par fixation. Ce saut est principalement dû à la longueur des mots (les mots courts sont plus facilement passés que les longs) (Brysbaert et Vitu, 1998), et, dans une moindre mesure, à leur fréquence ou à leur prédictibilité en fonction du contexte précédant (Rayner, 1998, 387). Ainsi, si nous fixons 85% des mots de contenus, seuls 35% des mots fonctionnels le sont (Carpenter and Just, 1983, cités par Rayner 1998, 375).

3. Les mouvements de régression : il en existe de deux types, qui sont en fait des saccades particulières. Le premier consiste à revenir au début de la ligne suivante lorsque la ligne en cours de lecture est terminée. Cette saccade est plus longue que les autres : environ 80 ms d'après Lévy-Schoen et O'Regan (1978, 463), ce qui fait que souvent, les lecteurs ont tendance à sous-estimer la distance et doivent effectuer une petite saccade supplémentaire pour se positionner correctement au début de la ligne (Rayner et Pollatsek, 1989, 114).

   Toutefois, le mouvement que l'on qualifie habituellement de régressif se manifeste chez le lecteur qui éprouve des difficultés face à un texte trop complexe pour une lecture courante ou, simplement, parce qu'il est fatigué ou distrait. Rayner (1998, 387) rapporte que 10 à 15% du temps de lecture est consacré à ces retours en arrière, certains étant très courts (quelques lettres seulement), d'autres beaucoup plus longs. Les premiers, souvent consécutifs à une longue saccade, s'expliqueraient par des erreurs dans la programmation du point de fixation (Vitu, McConkie and Zola, 1998, cités par

> Rayner 1998, 387), tandis que les seconds résulteraient d'un problème de compréhension. Dans ce cas, le lecteur peut revenir en arrière soit quand il rencontre un terme remettant en question l'interprétation qu'il a faite de la phrase, soit quand il veut se remémorer des éléments oubliés par la mémoire de travail.

Diverses études ont montré que certaines caractéristiques de ces trois types de mouvements (durée de fixation, longueur des saccades et de l'empan perceptuel ou nombre de fixations et de régressions) varient en fonction de l'expertise du lecteur, de l'attitude du lecteur vis-à-vis du texte et des caractéristiques du texte (typographie, difficulté, langue...) (Lévy-Schoen et O'Regan, 1978, 467).

Ainsi, depuis le début du siècle, on a cherché à mettre en relation les mouvements des yeux et le niveau du lecteur. Richaudeau (1969, 25) estime que la taille de l'empan fait le lecteur rapide, tandis que pour Alderson (2000, 56), ce serait plutôt la vitesse des fixations et des processus mis en œuvre durant ce temps de fixation. Il semblerait toutefois qu'une pratique régulière de la lecture influence simultanément chacune de ces caractéristiques, comme le rapporte Rayner *et al.* (2001, 46) :

> *First, beginning readers eye movements are quite different from those of skilled readers (Rayner, 1986). Beginning readers fixate virtually every word (and make more than one fixation on many words). Thus, their saccades are much shorter (around three letter spaces) than skilled readers'. Furthermore, their average fixation durations are much longer (between 300 and 400 ms) than skilled readers', and they regress much more frequently (so that up to 50% of their eye movements are regressions). Their perceptual span is also smaller than that of skilled readers.*

Dans le cadre de nos recherches en lisibilité, nous sommes davantage intéressé par un second type de variables : les particularités du texte qui agissent sur la nature des mouvements oculaires. En effet, puisque la majorité des auteurs s'accordent à dire que les mouvements des yeux reflètent bien l'évolution des activités cognitives du lecteur moment par moment (Just et Carpenter, 1980; Rayner, 1998; Rayner et Pollatsek, 1989), il paraît dès lors hautement plausible que là où l'oeil hésite, la tête achoppe, et que les variables qui ralentissent le mouvement des yeux prennent part à la difficulté du texte. Nous ne faisons ici que les esquisser, puisque le chapitre 3, qui décrit notamment les études empiriques justifiant l'emploi de certaines caractéristiques textuelles comme indices de la difficulté, va détailler ces diverses variables.

Tout d'abord, certaines études ont analysé l'effet de la typographie et la distance de lecture sur le mouvement des yeux. Pour Richaudeau (1969, 1), « le nombre de saccades est indépendant de la distance entre le lecteur et le livre », de même que le nombre de signes

appréhendés. Quant à la taille des caractères, Dehaene (2007, 37) explique que si une grande taille de police peut empêcher en partie la diminution de reconnaissance en région parafovéale, « plus un mot est écrit en gros caractères, plus ceux-ci prennent de la place sur la rétine et donc plus les lettres s'écartent vers la périphérie du champ visuel. ». Aussi, ces deux facteurs s'annulent, bien que Richaudeau (1969) rapporte un seuil minimal pour la taille des caractères, en-deçà duquel la fovea elle-même ne peut plus les reconnaître.

Ensuite, la fréquence des mots s'est révélée influencer directement le temps de fixation de ces mots (Rayner, 1998, 376). Ainsi, les mots peu fréquents sont fixés plus longtemps que les mots très fréquents (Just et Carpenter, 1980; Rayner *et al.*, 2001; Rayner et Pollatsek, 1989). De même, certains chercheurs (Rayner et Pollatsek, 1989, 158) ont montré une variation du temps de fixation en fonction de la place du mot dans la phrase, le verbe étant généralement fixé plus longuement, de même que les mots ambigus. À l'inverse, les éléments aisément prédictibles d'après le contexte qui les précèdent requièrent des fixations plus courtes.

Chez le lecteur d'une langue étrangère, on peut s'attendre à ce que ses mouvements oculaires se rapprochent de ceux d'un lecteur peu expérimenté. À notre connaissance, trop peu d'études se sont penchées sur la question, mais les rares qui existent s'accordent remarquablement. Ainsi, Oller (1972a, 314) et Oller et Tullius (1973) ont observé, d'une part, que la quantité de régressions chez les lecteurs de l'anglais comme L2 n'étaient pas significativement différents de celle des lecteurs natifs ; d'autre part, que si les non-natifs lisent en moyenne moins vite, à l'aide d'un empan plus réduit et en effectuant plus de fixations, la différence principale entre des lecteurs natifs et non-natifs de l'anglais porte sur la durée des fixations, nettement plus longue chez les seconds. Cette observation a été répliquée par Bernardt (1987, citée par Gaonac'h 2005), laquelle estime également que la principale différence entre un lecteur natif et un lecteur non natif est le temps consacré à chaque fixation oculaire.

Ces fixations particulièrement longues semblent nous indiquer que les lecteurs en L2 ont besoin de plus de temps pour effectuer les processus de traitement, et notamment au niveau de la reconnaissance lexicale. C'est pourquoi, dans la suite de cette section, nous allons nous attacher à décrire le processus de reconnaissance des mots, ainsi que les problèmes spécifiques qu'il pose à des lecteurs non-natifs.

La reconnaissance des mots

Après avoir éclairci la manière dont l'appareil visuel perçoit l'input, le second problème que nous envisageons est la façon dont ces éléments disparates que représentent des traits sur un support écrit sont reconnus comme un mot appartenant au lexique mental d'un indi-

vidu (Lupker, 2005, 39), une thématique désignée sous le terme de reconnaissance visuelle des mots. Comme l'a défendu notamment Dehaene (2007) (*cf.* Section 1.2.3), ce processus constitue l'aspect réellement spécifique de l'activité de lecture : il est donc normal qu'il ait fait l'objet d'un grand nombre d'études. Plusieurs modèles actuels de la lecture (Harm et Seidenberg, 2004; McClelland et Rumelhart, 1981; Seidenberg et McClelland, 1989), que nous envisageons par la suite, se limitent d'ailleurs à ce niveau de description.

Un premier objet de débat en reconnaissance des mots est le type d'information qui serait extrait durant la perception visuelle et la manière dont celle-ci est utilisée pour retrouver le mot du lexique mental auquel elle correspond. La première conception ayant prévalu est l'**approche globale**, qui remonte aussi loin que James Cattell (1886, cité par Larson, 2004), et qui postule que l'on reconnaît un mot comme un tout en se basant sur sa forme globale ou enveloppe. Cette notion d'enveloppe est définie plus en détail par Bouma (1971) comme un motif représenté par les lettres montantes (ex. « t » ou « d »), descendantes (ex. « p » ou « q ») ou neutres (ex. « n » ou « a ») qu'un mot contient. Cette enveloppe est définie pour un mot écrit en minuscules, mais elle disparaît évidemment dans sa version en majuscules.

Cette dernière propriété a été utilisée par divers chercheurs pour éprouver l'hypothèse de l'approche globale. Parmi ceux-ci, Smith (1969) a obtenu une réduction de la vitesse de lecture entre un même texte écrit en minuscules et en majuscule. Celle-ci est toutefois faible (entre 5% et 10%) et Smith l'explique plutôt par un problème de discrimination au niveau des lettres qu'au niveau de la forme globale.

Cette interprétation, qui remet en question l'approche globale, conduit Gough (1972) à postuler une alternative à celle-ci : la **lecture sérielle lettre à lettre**. Autrement dit, on lirait un mot de gauche à droite, en commençant par la première lettre, puis la deuxième, etc. L'avantage de ce modèle est qu'il explique le fait que l'on lise plus vite un mot court qu'un mot long, ce qui est davantage problématique pour l'approche globale. Par contre, il a assez vite été abandonné, car il n'explique par un effet très important en psycholinguistique : l'effet de supériorité du mot.

L'effet de supériorité du mot désigne le fait qu'un mot présenté durant un court instant est mieux identifié qu'une suite de lettres aléatoire de même longueur. Il a été détecté par Reicher (1969) à l'aide d'une expérience assez ingénieuse. Il s'agit de présenter très brièvement un mot de quatre lettres, tel que « WORD », suivi d'un masque de même taille composé de dièses, lequel est lui-même suivi d'une tâche de sélection entre deux lettres (ex. : K et D pour la dernière lettre), dont l'une était présente dans le mot entrevu et l'autre donne un mot qui existe également dans la langue (ici, « WORK »). Le fait de proposer des distracteurs crédibles permet de contrôler les stratégies de devinement. Par ailleurs, le fait de recourir à une tâche de discrimination entre deux lettres plutôt qu'à une tâche de rappel, permet aussi

de contrôler les biais liés à un oubli.

L'une des conséquences de cette découverte, est qu'il est vraisemblable que la lecture d'un mot s'opère plutôt au niveau de la lettre, mais de façon simultanée et **en parallèle**. Plus exactement, la reconnaissance des mots passerait par l'activation, via une série de traits perçus visuellement, de lettres abstraites, c'est-à-dire indépendantes de la typographie et de la casse (McClelland et Rumelhart, 1981). Une fois ces lettres reconnues, elles activeraient ensuite les représentations des mots en mémoire, permettant alors leur reconnaissance. Pour McClelland et Rumelhart (1981), l'effet de supériorité des mots proviendrait d'une rétroaction d'information depuis le niveau des représentations des mots vers le niveau des lettres. D'autres hypothèses ont été proposées par la suite (*cf.* Section 1.2.3) pour expliquer cet effet, mais l'hypothèse de la perception en parallèle reste aujourd'hui majoritairement acceptée dans la discipline.

Le second grand débat ayant animé le domaine de la reconnaissance des mots concerne la manière dont, une fois les caractères perçus, s'effectue l'accès au sens. Deux grandes conceptions se sont opposées à ce sujet : la **voie directe**, qui fait l'hypothèse d'un accès direct à partir de la forme graphique, et la **voie phonologique**, qui postule un accès via l'intermédiaire des représentations phonologiques, même lors de la lecture silencieuse.

La voie directe a longtemps été considérée comme l'unique moyen de lire couramment, car on considérait que le codage phonologique était plus coûteux en ressources mentales que l'accès direct. Cette première conception trouvait ses justifications dans les arguments suivants (Harm et Seidenberg, 2004, 664) :

– le lecteur, s'il passe systématiquement par la voie phonologique, va être confronté à une ambiguïté lorsqu'il rencontre un homophone et il devra alors faire appel à la voie directe. Or, comme les homophones sont généralement nombreux dans la majorité des langues, le passage généralisé par la voie directe semble plus efficace afin d'éviter une surcharge de calcul ;
– Smith (1971) a avancé que le fait de se reposer sur une série de règles de recodage des graphèmes vers les phonèmes est aléatoire en anglais, tant il y existe de cas particuliers. Il estime que la maîtrise de ce système de règles semble trop complexe à acquérir et que son efficacité paraît discutable, d'où il conclut que la voie directe est la seule viable ;
– enfin, Smith (1973, cité par Harm et Seidenberg, 2004, 664) en vient à la conclusion, à l'aide de concepts issus de la théorie de la communication de Shannon et Weaver (1949), que le passage par un double codage (orthographique $\rightarrow$ phonologique $\rightarrow$ orthographique) est trop lent pour expliquer la vitesse de lecture de bons lecteurs.

Ces différents arguments vont avoir une large influence sur le domaine durant près d'une

vingtaine d'années. La conception de Smith (1971) va constituer un fondement théorique pour l'approche globale (Rayner *et al.*, 2001), puisque apprendre à lire revient à intégrer les correspondances entre des formes orthographiques et leurs représentations sémantiques. Elle va également se répercuter dans les recherches en L2. Ainsi, Coste (1978, 29) défend l'accès direct comme objectif de la lecture en langue étrangère. De même, Cicurel (1991, 11) avance : « Dans la lecture silencieuse, le lecteur entraîné attribue directement un sens à la forme graphique en éliminant le stade de l'oralisation des phonèmes ».

Parallèlement à cette conception, quelques chercheurs, parmi lesquels Gough (1972), défendent une vision drastiquement opposée : la voie phonologique. Celle-ci postule que l'activation phonologique d'un mot précède nécessairement sa reconnaissance. Elle s'appuie notamment sur l'argument que l'homme apprend d'abord à parler avant de lire et que les mots seraient dès lors conservés en mémoire sous une forme orale, rendant leur activation obligatoire avant l'accès aux représentations sémantiques. Elle n'a toutefois pas fait florès, car elle a été assez vite supplantée par deux théories situées entre ces deux extrêmes.

Van Orden et ses collègues (Van Orden, 1987; Van Orden *et al.*, 1988, 1990) proposent ainsi la **théorie de la cohérence phonologique**. Ils reprennent l'axiome de la primauté de l'oral sur l'écrit dans le développement de l'homme et observent qu'en anglais, la correspondance entre l'orthographe et la sémantique est largement arbitraire, tandis que celle entre l'orthographe et la phonologie est davantage systématique. Par conséquent, la première devrait être plus ardue à apprendre que la seconde (Harm et Seidenberg, 2004, 665), et comme l'apprenti-lecteur a déjà intégré la relation phonologie $\rightarrow$ sémantique de nombreux mots, il va préférer effectuer le lien entre la forme orthographique et la représentation sémantique via la forme phonologique. Autrement dit, « le système orthographique vient dont se greffer de façon parasite sur le système du langage parlé déjà existant » (Ferrand, 2007, 201), ce qui ne l'empêche pas de créer peu à peu des connexions directes vers le niveau sémantique.

Une seconde théorie cherchant à réunir plus explicitement encore ces deux types de codage est la **théorie des deux routes** (Coltheart, 1978; Seidenberg et McClelland, 1989), qui est détaillée à la Section 1.2.3. Brièvement, on peut dire que la voie directe et la voie par activation phonologique représentent deux accès possibles au sens. Pour un bon lecteur, c'est l'accès direct qui est privilégié, car l'activation phonologique est plus lente. La route phonologique, qui est donc optionnelle, peut néanmoins être utilisée dans certaines circonstances, telles que la lecture d'un mot inconnu ou d'un non-mot. Selon cette conception, la reconnaissance d'un mot s'opère donc via l'une ou l'autre de ces deux voies. Harm et Seidenberg (2004) ont toutefois révisé cette position, défendant le fait que ces deux routes pourraient bien coopérer, plutôt que d'opérer de manière indépendante.

Par ailleurs, divers travaux en neuropsychologie de la lecture ont montré que ces deux voies

existent bel et bien et qu'elles font appel à des réseaux différents du cerveau (Jobard *et al.*, 2003). Dehaene (2007) rapporte de plus l'existence d'une région particulière, connue sous le nom latin de planum temporale, qui serait le lieu de la compatibilité entre les lettres et les sons. Par conséquent, il jouerait un rôle essentiel dans l'apprentissage de la lecture, puisque « lorsque nous ânonnons notre b.a-ba, le planum temporale apprend progressivement à reconnaître les correspondances entre la sonorité et l'apparence des lettres. À l'âge adulte, ces liens entre graphèmes et phonèmes s'automatisent et prennent la forme d'un véritable réflexe de conversion des lettres vers les sons » (Dehaene, 2007, 152).

En résumé, il apparaît donc que le décodage des mots s'effectuerait au niveau de la lettre, de façon parallèle, et qu'il mettrait en jeu non seulement les correspondances entre des codes orthographiques et les représentations sémantiques, mais également celles qu'entretiennent les représentations phonologiques avec le lexique, même dans le cas de la lecture silencieuse.

Il semblerait que ces conclusions restent globalement valides pour la lecture dans une L2, si ce n'est que la présence d'un second (voire plus) ensemble de représentations, liées à la L2, rend la situation plus complexe. De nouvelles questions émergent dès lors de ce constat, dont la plus importante est sans doute de déterminer si, chez un bilingue, le processus de reconnaissance des mots n'implique que l'activation de la représentation du mot dans la langue cible (**accès sélectif au lexique**) ou si l'ensemble des mots proches, y compris ceux de l'autre langue, sont activés (**accès non sélectif**) (Lemhöfer *et al.*, 2008, 12).

À ce propos, la majorité des études [18] auraient tranché en faveur d'une interaction entre les deux langues lors de la reconnaissance de mots dans l'une des langues, notamment suite à des activations conjointes lors du traitement de congénères ou de faux amis. Nous reviendrons plus en détail sur cette question dans le Chapitre 3. Ces travaux ont également débouché sur un modèle de la reconnaissance des mots dans le cadre bilingue, le Bilingual Interactive Activation (BIA) de Van Heuven *et al.* (1998), qui est décrit à la Section 1.2.3.

Enfin, le dernier aspect des processus de reconnaissance des mots qui nous intéresse ici est son influence sur la compréhension des textes. En effet, dans le contexte d'une étude sur la lisibilité des textes, on peut se demander en quoi ces problématiques de décodage importent lorsqu'il s'agit, comme dans le cas de la lisibilité, d'évaluer avant tout la compréhension que des lecteurs se font d'un texte. La réponse à cette interrogation est liée à deux concepts introduits par LaBerge et Samuels (1974) : l'automatisation et l'attention.

L'automatisation de la reconnaissance des mots d'un texte est atteinte lorsque le lecteur n'est pas conscient que ce processus est en train de s'opérer et que très peu de capacités de traitement sont mobilisées pour ce faire (Grabe, 1991, 379). Quant à l'attention, LaBerge

18. Pour un inventaire plus détaillé de celles-ci, consulter Dijkstra et Van Heuven (2002).

et Samuels (1974) estiment qu'il s'agit d'une ressource mentale assez limitée (elle ne peut s'appliquer qu'à une opération à la fois), mais nécessaire à la lecture. Dès lors, confronté à cette activité aux multiples composantes, le lecteur ne dispose que de deux possibilités : soit il automatise certaines tâches, soit il opère des changements d'attention, se concentrant tantôt sur le décodage des mots, tantôt sur la compréhension des idées. Si le lecteur faible recourt à cette dernière stratégie, qui lui demande du temps, le bon lecteur aura automatisé la reconnaissance des mots et pourra alors se concentrer sur les processus de haut niveau.

Cette notion d'automatisation, que les auteurs fondent en rapportant une série impressionnante d'expériences à divers niveaux de traitement, a connu un beau succès dans les recherches en lecture et notamment celles sur la L2. En effet, le concept d'attention apparaît fécond pour expliquer certaines difficultés des lecteurs en L2 dont les ressources limitées seraient d'autant plus sollicitées que le décodage ne peut se reposer sur des connaissances linguistiques préalables. Samuels et Kamil (1984, 197) envisagent d'ailleurs la situation où un apprenant doit recourir au basculement d'attention :

> *This alternative switching of attention from decoding to comprehension is similar to the strategy used by beginning students of a foreign language who first work their way through a novel written in a foreign language by translating all the difficult words and then rereading the text again in order to understand it.*

Plus généralement, l'idée que la qualité du décodage influence le niveau de compréhension a été soulignée par plusieurs auteurs. Crain et Shankweiler (1988) observent ainsi que les lecteurs faibles commettent plus d'erreurs de compréhension sur des relatives et que, parallèlement, ils obtiennent un moins bon taux de rappel immédiat que les lecteurs avancés. Il y a donc bien une différence dans la qualité de la mémorisation, qui va de pair avec un défaut de compréhension. Or, ces différences de mémorisation ne se retrouvent pas dans d'autres types de tâches (telles que la reconnaissance de visage). On doit conclure à une infériorité de leur mémoire uniquement dans les processus linguistiques, ce qui pourrait s'expliquer par le fait que plus d'efforts aient été consentis au niveau du décodage. Par ailleurs, il existe plusieurs études qui confirment l'existence d'une corrélation entre la vitesse de décodage et la qualité de compréhension (Crain et Shankweiler, 1988; Perfetti et Hogaboam, 1975; Rayner et Pollatsek, 1989).

Il est donc aujourd'hui largement reconnu que bien savoir lire requiert d'automatiser l'étape de décodage, que ce soit en L1 comme en L2 (Grabe, 1991, 379). C'est pourquoi il apparaît que la prise en compte des éléments textuels influençant ce processus est susceptible de se révéler intéressante en lisibilité, même quand c'est la compréhension qui est visée avant tout.

La question de la compréhension

La dernière grande composante de la lecture dont nous allons traiter inclut les processus de haut niveau participant à la compréhension du texte. Comme d'autres processus mentaux complexes, la notion de compréhension est ardue à formaliser, notamment parce qu'elle fait intervenir un grand nombre de sous-processus interactifs, qui rendent son observation ou sa manipulation délicate. Elle a néanmoins attiré l'intérêt des chercheurs depuis longtemps. Henderson (1903, cité par Venezky, 1984, 15) avait déjà montré la prédominance d'une organisation conceptuelle des contenus du texte en mémoire par rapport à une simple conservation de l'image graphique en mesurant la trace du processus de lecture au moyen d'une reproduction écrite des souvenirs de lecture. Par la suite, Bartlett (1932, cité par Venezky, 1984, 15) confirme cette conclusion. Il observe des regroupements d'idées, des simplifications et des modifications, dues à l'influence des conceptions préalables du sujet. Il va même jusqu'à considérer que ce qui est conservé en mémoire sur le long terme est un « schème », faisant ainsi office de précurseur de l'approche récente de la théorie des schèmes dont nous parlerons par la suite.

Ces premiers travaux mettent donc déjà en évidence une caractéristique essentielle de la compréhension écrite : elle consiste en l'extraction d'informations depuis le support textuel et en leur organisation au sein d'un réseau préexistant. Plus précisément, Dubois (1976, citée par Vigner, 1979, 37) définit la compréhension comme « l'ensemble des activités qui permettent l'analyse des informations reçues en termes de classes d'équivalences fonctionnelles, c'est-à-dire l'ensemble des activités de mise en relation d'informations nouvelles avec des données antérieurement acquises et stockées en mémoire à long terme ». Cette définition soulève donc deux problématiques : comment s'opère l'analyse des informations du texte et comment celles-ci sont-elles stockées en mémoire ? Nous allons revenir brièvement sur ces deux problématiques dans la suite de cette section.

Auparavant, soulignons encore qu'il est de plus en plus accepté que les processus de la compréhension écrite ne sont pas spécifiques à la lecture, à la différence des processus de reconnaissance des mots (Dehaene, 2007; Hoover et Gough, 1990; Hoover et Tunmer, 1993; Rayner *et al.*, 2001; Smiley *et al.*, 1977). D'après Dehaene (2007, 147), les régions activées 250ms après la présentation d'un stimulus, et qui correspondent aux aires du cerveau utilisées pour la compréhension, s'activent de la même façon suite à un stimulus oral ou écrit. La compréhension serait donc une activité cognitive dépassant de loin la question de la lecture, qui doit cependant être distinguée de la résolution de problèmes. Comme nous l'avons dit précédemment, on peut y distinguer deux étapes : l'extraction du sens, qui conduit à la construction d'une structure du texte en mémoire, et la mise en rapport de celle-ci avec le réseau de connaissances et de représentations du lecteur.

En ce qui concerne la formation de la structure textuelle, l'un des modèles les plus influents est celui de Kintsch et van Dijk (1978), qui a connu plusieurs révisions (Kintch, 1998; Kintsch, 1988). Il postule qu'une fois les mots reconnus et leurs sens activés, ceux-ci sont combinés en un ensemble de propositions à l'aide de divers indices tels que les relations syntaxiques et divers marqueurs de la cohérence. Cette étape aboutit à la construction de ce que l'on appelle la **microstructure** du texte, c'est-à-dire un réseau de propositions reliées entre elles par des divers liens logiques, tels que la co-référence. Les auteurs soulignent que la construction d'une telle structure requiert bien souvent la production d'inférences, afin de relier entre elles des propositions non directement connectées dans le texte. De nombreux détails supplémentaires concernant cette étape sont décrits dans les Chapitres 2 et 3.

Cependant, la production d'une microstructure dans l'esprit du lecteur ne suffit pas. En effet, confronté à un texte plus long, il est nécessaire d'en articuler les différentes sections en une structure de signification globale construite sur la base des propositions du texte : la **macrostructure**. Cette construction s'opèrerait via différentes opérations effectuées sur les propositions de la microstructure : il s'agit notamment de l'identification des idées principales, mais aussi de généralisations, de suppressions, comme l'indiquent Kintsch et van Dijk (1978, 380) :

> *Propositions that are not generalized are deleted if they are irrelevant ; if they are relevant, they may become macropropositions.*

La construction de la macrostructure peut être guidée par des modèles préfabriqués, qui facilitent l'articulation des propositions : les schémas, dont nous avons déjà parlé précédemment. Ils correspondent à « un groupement structuré de connaissances qui représentent un concept particulier » (Cornaire, 1999, 25) et sont organisés au sein d'une structure hiérarchisée comprenant l'ensemble des schémas et se modifiant sous l'apport de nouvelles informations. Anderson et Pichert (1978, 2) envisagent deux façons dont les schémas peuvent influencer la sélection des informations lors de la lecture : soit le schéma met en évidence les éléments importants du texte, soit il contient des « slots » qui doivent être remplis et servent ainsi à sélectionner les informations.

Plus précis dans leur analyse, Anderson et Pearson (1984) se positionnent en faveur de la seconde optique. Ils estiment qu'au cours de la lecture, le lecteur rencontre divers mots susceptibles de remplir les « slots » de schémas qu'il a mémorisés et que le plus probable parmi ceux-ci va être activé. Le schéma retenu permet d'organiser les propositions extraites du texte via une série d'inférences qui peuvent servir (1) à compléter un « slot » du schéma activé avec un mot issu du texte ; (2) dans le cas où aucun élément du texte ne permet de remplir ce « slot », à utiliser des connaissances préalables pour lui attribuer une valeur par défaut ; (3) tirer des conclusions qui ne sont pas présentes dans le texte.

Les schémas expliquent donc à la fois comment les propositions sont organisées en une macrostructure, mais aussi comment les connaissances déjà emmagasinées en mémoire jouent un rôle dans le processus d'interprétation des nouvelles informations extraites du texte, et dans l'intégration de celles-ci au sein la mémoire sémantique (Anderson et Pearson, 1984, 256). On rejoint ici la seconde partie de l'étape de compréhension, durant laquelle les informations extraites du texte sont mises en relation avec des connaissances préalables et intégrées à la structure mémorielle préexistante. C'est au terme de cette seconde étape que l'on peut dire que le texte a été compris. Les informations mémorisées pourront alors être récupérées, notamment dans le cadre d'un test de compréhension.

Pour en conclure avec la compréhension, soulignons que le concept de schéma sera repris par Rosenblatt (1978, citée par Tracey et Morrow 2006, 55) afin d'expliquer la spécificité de chaque lecture, puisque chaque lecteur recourt à des schémas résultant de sa propre expérience. Par conséquent, cette vision a des conséquences évidentes dans le cadre L2, où *a fortiori*, les lecteurs issus d'une autre culture auront plus de difficulté à activer un schéma adapté au texte qu'ils sont en train de lire. D'ailleurs, diverses études (Cabello, 1984; Steffensen *et al.*, 1979) ont montré que la compréhension de textes rédigés selon des schémas spécifiques à sa propre culture était supérieure à celle de textes utilisant des schémas étrangers.

Conclusion

Le parcours de ces trois composantes principales du processus de lecture a permis de se forger une idée plus précise des principaux processus mis en oeuvre lors de la lecture. Au-delà de l'intérêt qu'il y a à poser ce cadre avant d'aborder le domaine plus spécifique de la lisibilité, ce qui sera fait dans le chapitre 2, ce détour par les études de psychologie cognitive a permis de formuler une hypothèse essentielle pour notre approche.

Ainsi, la présentation d'une série des résultats d'expériences empiriques sur les mouvements oculaires et la reconnaissance des mots a mis en évidence l'existence de caractéristiques textuelles influençant le processus de lecture. Nous avons ainsi souligné l'effet de la longueur des mots sur les points de fixation, ainsi que le rôle que la fréquence des mots ou leur place dans la phrase joue sur le temps de fixation. En ralentissant potentiellement la reconnaissance des mots, ces caractéristiques peuvent influencer le processus de compréhension. En effet, comme nous l'avons vu, les ressources attentionnelles nécessaires pour extraire des propositions du texte et les articuler au sein d'une structure cohérente sont limitées et le fait de consacrer davantage de temps aux processus de niveau inférieur allonge le temps de traitement. Or, au vu des limites de la mémoire à court terme (Baddeley, 2003; Miller, 1956), cet accroissement du temps de traitement augmente également la possibilité d'oublier une

partie de la structure textuelle mémorisée précédemment, ce qui pénalise la compréhension.

Il s'agit là d'une hypothèse essentielle pour notre thèse, qui explique pourquoi on peut s'intéresser à des caractéristiques textuelles de surface pour chercher à prédire automatiquement la difficulté qu'un apprenant de FLE éprouve pour lire et comprendre un texte donné. Se basant sur ce postulat, il convient évidemment d'étendre l'ensemble des caractéristiques envisagées, prenant notamment aussi en compte le nombre de sens d'un terme, l'âge d'acquisition des mots, la redondance, la concrétude des mots, la longueur de la phrase, etc. Certaines de ces variables sont bien connues en lisibilité, mais d'autres n'ont pas encore — à notre connaissance — fait l'objet d'étude dans ce cadre particulier. Nous reviendrons toutefois plus en détail sur cette question dans le chapitre 3. La section suivante décrit quant à elle les principaux modèles qui ont tenté d'articuler les différentes composantes de la lecture décrites dans ce point.

1.2.3 Les théories de la lecture

Comme il l'a déjà été dit précédemment, la lecture, en langue maternelle comme en langue seconde, reste réfractaire aux tentatives de description. Comme d'autres activités cognitives, elle présente ceci de délicat que l'essentiel se passe à l'intérieur. On a longtemps imaginé ne l'étudier que comme une boîte noire (Richaudeau, 1969, 30) : à l'entrée, il est possible de contrôler l'input (le texte) et d'observer les mouvements des yeux ; à la sortie, de tester les souvenirs, c'est-à-dire le produit [19] de la lecture, ce qui a conduit aux deux principales optiques méthodologiques du domaine, selon que les chercheurs ont privilégié l'analyse des processus (via l'observation des mouvements oculaires) ou du produit de la lecture.

Ainsi, inspirées par le structuralisme — qui, en psychologie, cherche à expliquer la structure de la pensée sur la base de l'étude de la perception (Tracey et Morrow, 2006, 22) —, les premières approches des processus de lecture se sont concentrées sur la perception visuelle. Elles datent de la fin du 19e siècle et sont dues à Wundt, Javal, Cattell, Huey, etc... (cf. Venezky, 1984). Toutefois, à partir de 1910, cet attrait vers les processus cognitifs est peu à peu négligé sous l'effet de considérations pédagogiques telles que la conception de tests de compréhension à la lecture et l'impulsion d'un nouveau de courant de pensée en psychologie, le behaviourisme. Tracey et Morrow (2006, 32) résume bien ce tournant méthodologique :

> *Behaviorism changed the depiction of reading from one of perceptual processing to one of reading as a behavior composed of isolated skills, each of which*

19. Cette distinction entre *processus* et le *produit* de ce processus est empruntée à Alderson (2000, 3), qui précise que le processus est silencieux et interne tandis que le produit est sujet à la variation personnelle sans que l'on sache si cette variation s'ancre dans le processus en lui-même ou dans les compétences périphériques des lecteurs.

could be reinforced to increase student achievement.

Dès lors, l'analyse du produit connu une large popularité dans la première moitié du XX^{e} siècle, notamment avec les travaux en lisibilité.

Parallèlement se développe une autre théorie de l'apprentissage, le constructivisme, qui insiste sur le fait que l'acquisition de connaissances est le résultat d'une construction active de la part des individus. Elle introduit notamment la notion de schèmes, définie par Bartlett en 1932 et dont nous avons déjà parlé à la Section 1.2.2. Par ailleurs, le constructivisme a également mis en avant l'usage de stratégies métacognitives par les apprenants, lesquelles ont fait l'objet de nombreuses études postérieures [20].

Il faut attendre les années 60 pour que renaisse un vif intérêt à l'égard des processus cognitifs de la lecture, sous l'influence de la psychologie cognitive. Tardif (1990, 9) décrit cette dernière comme une branche de la psychologie qui « concerne essentiellement l'étude scientifique des phénomènes mentaux [et qui] s'intéresse au traitement de l'information par le cerveau au niveau de la réception, de l'interprétation et de la réutilisation ». Cette période a été marquée par l'emploi de méthodes d'investigation plus variées, que l'on regroupe sous l'appellation « chronométrie mentale » (Ferrand, 2007). Celles-ci comprennent : la tâche de décision lexicale, où un sujet doit décider le plus rapidement possible si la suite de lettres qui lui sont présentées appartient ou non au lexique de sa langue (Rubenstein et *al.*, 1970, cités par Ferrand 2007, 33) ; la lecture orale ; la tâche d'identification perceptive (due à Howes et Solomon 1951) qui consiste à identifier un mot dégradé visuellement ; la catégorisation sémantique, initialement utilisée par Meyer and Ruddy (1975, cités par Ferrand 2007, 37), où le sujet doit décider si un mot appartient ou non à une classe sémantique donnée ; la technique d'amorçage, où un premier stimulus (l'amorce) est présenté brièvement avant un second (la cible) afin de déterminer si l'amorce produit un effet facilitateur ou inhibiteur sur le traitement de la cible.

À l'aide de ces techniques et des résultats expérimentaux qui découlent de leur usage [21], divers modèles de la lecture vont être élaborés. Ces derniers se veulent une description détaillée et organisée de l'ensemble des processus mentaux mis en œuvre par le sujet lors de la lecture, depuis la perception visuelle jusqu'à la mise en mémoire des informations extraites du texte. De plus, comme toute théorie scientifique digne de ce nom, ils doivent être

20. Le lecteur intéressé peut consulter la synthèse proposée par Tracey et Morrow (2006, 47-75).

21. Bien que ce ne soit pas notre propos ici, il convient de sensibiliser notre lecteur au fait que les résultats expérimentaux présentés dans ce chapitre, ainsi que les modèles qui en découlent ont été influencés par la nature de ces méthodes d'investigation. Ainsi, Samuels et Kamil (1984, 189) isolent au sein de ces techniques d'expérimentation quatre variables susceptibles de modifier les résultats : « These are the age and skill of the experimental subjects, the tasks which the subjects are asked to perform, the materials which are used, and the context (e.g., classroom, laboratory, type of school, etc.) which surrounds the study ». Pour un exposé plus détaillé des biais particuliers à chacune de ces méthodes, consulter Ferrand (2007, 29-43).

falsifiables et explicatifs (cf. Platt, 1964). C'est-à-dire qu'un modèle, qui consiste en une série d'assertions, doit expliquer l'ensemble des données empiriques récoltées jusque-là, mais doit aussi s'avérer suffisamment précis pour autoriser des prédictions (en effet, un modèle très vague explique tout, mais ne nous apprend rien de neuf). Ces hypothèses peuvent alors être testées à l'aide d'expérimentations. Si les prédictions sont vérifiées par ces expériences, le modèle est considéré comme vrai jusqu'à preuve du contraire [22] ; par contre, si elles se révèlent fausses, le modèle doit être raffiné afin de prendre en compte ces nouvelles observations. Ainsi, la connaissance se construit de manière incrémentale et les modèles deviennent de plus en plus précis.

Il convient toutefois de rappeler que, suite aux biais inhérents aux méthodes d'investigations employées en psychologie cognitive de la lecture, les théories présentées dans cette section doivent être considérés comme une description adéquate du processus de lecture sous une série de conditions particulières, et non valables pour l'ensemble des situations de lecture (Samuels et Kamil, 1984, 190). C'est d'autant plus vrai que chacun d'eux repose sur un ensemble de connaissances et de postulats qui furent souvent mis à mal par des conceptions postérieures.

Parmi ces postulats, on peut faire ressortir deux axes, en fonction desquels les modèles peuvent être classés : la nature sérielle ou parallèle des processus et la manière dont le lecteur exploite ses diverses connaissances. Ainsi, influencées par les théories psychologiques du traitement de l'information, dont le modèle de Atkinson et Shiffrin (1968) constitue l'exemple le plus célèbre, les constructions théoriques des années septante sont **séquentielles** : elles se représentent la lecture comme une série d'étapes se succédant sans retour en arrière possible et sans interaction. Par opposition, les modèles postérieurs sont **non séquentiels** : ils envisagent l'action simultanée de plusieurs sources d'informations.

Quant à l'emploi des connaissances préalables, certains modèles ont limité leur rôle à l'analyse des données extraites du texte au moyen de processus de niveau inférieur tels que la reconnaissance des mots, tandis que d'autres ont mis en évidence leur action au niveau des processus de niveau supérieur tels que l'organisation des idées issues des phrases en un modèle de compréhension du texte, la sélection des informations essentielles ou encore la mise en relation de ces informations avec des connaissances antérieures (Grabe, 2005, 52) [23].

Sur la base de ces deux dimensions, on peut regrouper les modèles des années 60 à 80 en

22. Pour une explication logique de l'impossibilité de prouver une théorie, nous renvoyons le lecteur au résumé de Gough (1984, 226-227).

23. En réalité, il existe évidemment d'autres axes permettant d'ordonner ces modèles. De Beaugrande (1981) en a isolés 16 parmi lesquels le type de stockage mémoriel, le niveau d'automatisation, la profondeur du traitement, etc. Il explicite également la différence entre le type d'organisation des différentes composantes (modularité / interactivité) et la nature du traitement (sériel ou parallèle).

trois catégories que nous décrivons ci-dessous : les modèles **ascendants** (« bottom-up »), où la compréhension se fait des graphèmes au sens, les modèles **descendants** (« top-down »), où le lecteur prédit le sens du texte à partir d'informations textuelles, mais surtout de connaissances qu'il possède en lui, et enfin, les modèles **interactifs**, qui considèrent que « [the] skills at all levels are interactively available to process and interpret the text » (Grabe, 1998, 59).

À partir des années 80, la psychologie cognitive a connu, à l'instar de la linguistique, une formalisation croissante à l'aide d'outils issus de la mathématique et de l'informatique, l'objectif étant de simuler divers comportements humains à l'aide de modèles implémentables. Ce nouveau paradigme présente certains avantages par rapport aux modélisations classiques qui reposent sur l'analyse du comportement de lecteurs normaux ou dyslexiques. D'abord, il oblige les chercheurs à être davantage explicites vis-à-vis de leurs hypothèses, qui sont dès lors testables quant à leur cohérence interne et plus aisément falsifiables. Le niveau élevé de formalisation facilite également la comparaison entre deux modèles, ce qui n'était guère possible auparavant. De plus, ces simulations « peuvent prédire des effets cachés ou trop complexes pour être appréhendés par l'être humain » (Ferrand, 2007, 48). Enfin, elles permettent de simuler de façon plus réaliste des dégradations consécutive à des dommages cérébraux, puisque certaines performances diminuent, mais que la totalité de la connaissance n'est pas perdue (Plaut, 2005, 25; Barber et Kutas, 2007, 100) .

Au sein de cette nouvelle manière d'aborder la lecture, on peut distinguer deux approches sur lesquels nous reviendrons par la suite : le **modèle à deux routes**, héritier de la philosophie modulaire qui régnait dans les années 70 et 80[24] (Coltheart *et al.*, 1993, 2001) ; et les **modèles connexionistes**, basés sur une correspondance avec l'architecture du cerveau (McClelland et Rumelhart, 1981; Rumelhart et McClelland, 1982; Rumelhart *et al.*, 1986b; Seidenberg et McClelland, 1989).

Enfin, depuis les années 70, des avancées technologiques telles que l'imagerie cérébrale ont ouvert le champ de la lecture à une troisième voie moderne : **la neuroscience**. À la différence de la perspective cognitive dont les racines s'ancrent dans la linguistique, les neurosciences approchent la cognition selon un point de vue issu de la biologie. Elles permettent d'entrouvrir cette boîte noire à la fine mécanique que constitue notre cerveau et révèlent notamment que le cerveau est loin d'être aussi flexible que ne l'ont pensé les chercheurs en sciences humaines (Dehaene, 2007, 26). Nous envisageons ces différentes approches dans les sections qui suivent.

24. Fiset et Arguin (2002, 181-182) rappellent que cette philosophie « postule que le cerveau est constitué de modules indépendants qui accomplissent leur travail de manière séquentielle, comme dans une chaîne de montage ».

Les modèles ascendants

Les modèles ascendants sont sémasiologiques : ils vont du signe au sens, considérant que la signification d'un texte se construit sur la base des données visuelles, en passant au travers de niveaux de traitements de plus en plus élevés, sans retour en arrière possible. En cela, ils constituent des représentants typiques de la conception séquentielle du traitement de l'information, que décrit bien Lupker (2005, 40) :

> *The human information processing system involves a series of processing stages that work in a serial, nonoverlapping fashion. Information only flows one way, that is, forward, through the system and, further, each stage is essentially completed before the next begins. The term « thresholded » is used to refer to the assumption that each stage is ready to pass information on to the next stage only when the activation at the initial stage reachers a threshold.*

Ils se caractérisent également par une focalisation sur la reconnaissance des mots, au détriment des mécanismes de la compréhension, ainsi que par une dépendance vis-à-vis du niveau oral : ils supposent que la reconnaissance des mots requiert le passage par un code oral (Cziko, 1980, 101) [25].

Les modèles les plus représentatifs de ce courant ont été développés par Gough (1972) et par LaBerge et Samuels (1974) [26]. L'article influent de Gough (1972) est considéré comme la première description globale des processus de lecture qui s'inspire de la théorie du traitement de l'information. Gough y décompose la lecture en une série d'étapes de traitement ascendantes (*cf.* Figure 1.1) qui s'amorcent dans la perception visuelle : les signes graphiques sont perçus par l'oeil et persistent un bref moment sous la forme d'une image visuelle que Gough (1972, 332) nomme l'« icon ». Celle-ci, constituée de traits, d'angles, de courbes, etc. est utilisée par le « scanner » pour reconnaître une à une les lettres, lesquelles sont ensuite transposées par le « decoder » en une chaîne de « phonèmes systématiques », à l'aide d'une série de règles de correspondances entre graphèmes et phonèmes. Cette représentation abstraite permet alors au lecteur d'accéder à son lexique mental et d'en extraire le sens du mot. Une fois les mots reconnus, ils sont entreposés en mémoire primaire afin d'y être organisés en unités plus larges à l'aide d'un mécanisme assez vague que Gough (1972, 341) nomme « Merlin ». Enfin, une fois comprises, les phrases sont transférées dans le PWSGWTAU [27].

25. Bien que, comme nous y reviendrons par la suite, LaBerge et Samuels (1974, 313) tiennent compte de la possibilité de passer directement de la forme graphique au sens, notamment dans le cas d'homophones.

26. Il faut cependant noter qu'ils ne sont pas spécifiques de la période cognitive, puisque Alderson (2000, 17) considère que : « this approach was typically associated with behaviourism in the 1940s and 1950s, and with "phonics" approaches to the teaching of reading that argue that children need to learn to recognize letters before they can read words (...) ».

27. Pour les lecteurs curieux, précisons que cet acronyme signifie « the Place Where Sentences Go When They Are Understood » Gough (1972, 341).

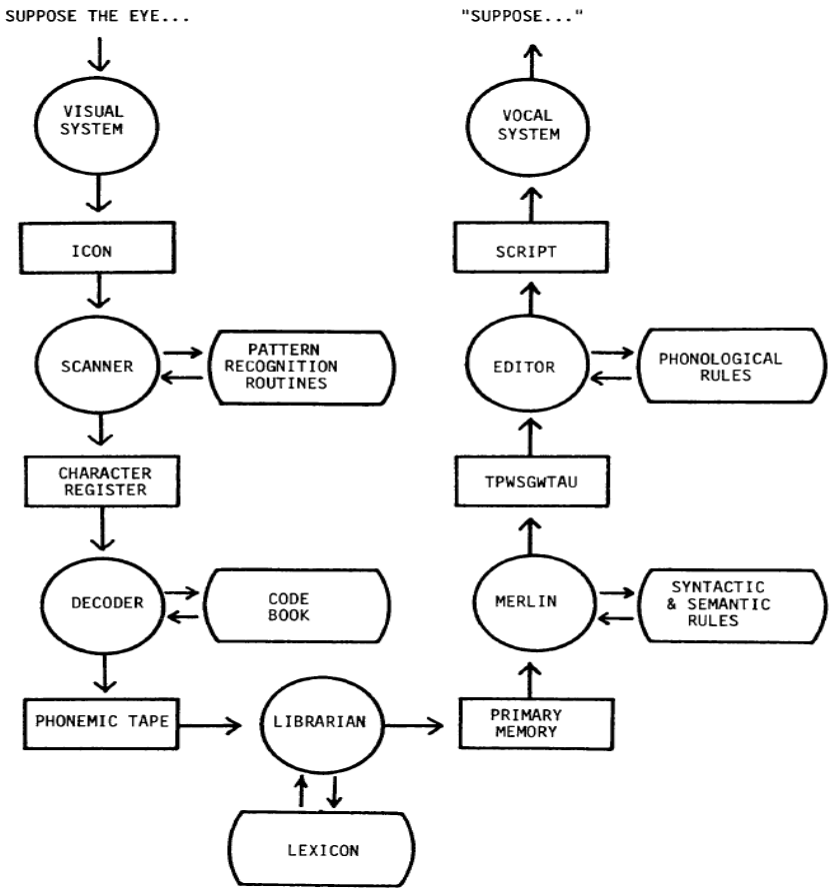


Figure 1.1 – Modèle de Gough (1972)

De nombreuses critiques ont été adressées au modèle de Gough (Brewer, 1972; Cornaire, 1985; Rayner et Pollatsek, 1989; Rumelhart, 1977). Elles s'appliquent notamment à sa conception de la lecture lettre-à-lettre, qui suggère qu'un mot doit être reconnu plus lentement qu'une lettre, alors que Urquhart et Weir (1998, 41) rapportent que des expériences ont prouvé le contraire. De même, le passage obligé par une forme phonologique n'a pas manqué de rencontrer des objections, puisqu'il ne permet pas d'expliquer comment un lecteur distingue des homophones (Brewer, 1972, 362).

Quant au modèle de LaBerge et Samuels (1974), son apport principal est la notion d'attention, quenous avons déjà présenté précédemment (*cf.* Section 1.2.2). On peut en conclure avec les modèles ascendants en signalant qu'ils n'ont pas eu autant de succès dans les études sur la lecture en français langue étrangère que les modèles descendants, que nous présentons maintenant.

Les modèles descendants

Les modèles descendants sont de nature onomasiologique : ils vont du sens au signe, c'est-à-dire que ce sont les processus cognitifs de niveaux supérieurs participant à l'élaboration du sens qui gèrent le contrôle du flux d'information à tous les niveaux (Rayner et Pollatsek, 1989, 461). Les modèles les plus connus dans cette catégorie sont dus à Goodman (1967) [28], Smith (1971) et, dans une moindre mesure Levin et Kaplan (1970). La littérature tend souvent à les opposer aux modèles ascendants quant à l'orientation du flux d'information, mais il semblerait que l'opposition entre ces deux approches soit mieux exprimée sous les termes « text-driven » et « reader-driven » (Urquhart et Weir, 1998, 42), parce que la caractéristique principale des modèles descendants tient au rôle prépondérant qu'y jouent les attentes du lecteur (ou hypothèses).

Cette conception repose sur le raisonnement suivant. Partant de l'hypothèse, proposée par Goodman (1970, cité par Rayner et Pollatsek, 1989) et Smith (1971), et fondée sur la théorie de la communication de Shannon et Weaver (1949), selon laquelle il existe des goulots d'étranglement dans le processus de lecture, tels que la quantité d'information qui peut être acquise en une fixation (Smith, 1971, 16) ou le nombre d'éléments que peut contenir la mémoire à court terme (Miller, 1956; Smith, 1971), à laquelle s'ajoute le présupposé que l'objectif du lecteur est d'extraire l'information du texte le plus rapidement et le plus efficacement possible, les auteurs déduisent que le lecteur, pour atteindre ce but, doit outrepasser les limites de sa perception visuelle et de sa mémoire en adaptant sa stratégie.

Pour ce faire, il émet des hypothèses sur la suite du texte (c'est-à-dire généralement le mot suivant) en s'appuyant sur ses connaissances préalables et, si elles sont disponibles, sur des informations contextuelles en provenance du texte. Ces hypothèses sont ensuite vérifiées à l'aide d'un nombre minimal d'indices (orthographiques, syntaxiques et sémantiques) issus du texte. Dès lors, un bon lecteur est considéré comme capable de tirer profit de la redondance inhérente à la langue, pour reconstruire le sens du message sans devoir traiter la totalité du matériel graphique (Smith, 1971, 220). Il pratique une lecture sélective (Goodman, 1967, 8; Smith, 1971, 94) , active et orientée par le contexte. Par ailleurs, il doit percevoir de manière globale et directe — c'est-à-dire sans passage par une forme orale ou phonologique — les mots, puisqu'au vu des limitations de sa mémoire à court terme, la lecture lettre à lettre ne permettrait pas d'expliquer la vitesse atteinte par les lecteurs expérimentés (Smith, 1971, 221).

28. Bien que Goodman n'ait jamais accepté cette catégorisation et que son modèle présente certaines caractéristiques des modèles interactifs, les spécialistes de la lecture (Cornaire, 1999; Cziko, 1980; Rayner et Pollatsek, 1989; Urquhart et Weir, 1998) s'accordent généralement à le classer parmi les modèles descendants, dont le représentant le plus typique reste toutefois Smith (1971).

Le paradigme descendant et l'approche globale de la lecture ont connu un écrasant succès, aussi bien auprès des éducateurs que des chercheurs explorant la lecture en langue étrangère, comme le révèle le recensement bibliographique effectué par Bernhardt (1991, 22-25) entre 1974-1988. Citons les travaux de Coady (1979) ou de Hosenfeld (1984), qui se fondent sur cette vision psycholinguistique de la lecture ou ceux de Rigg (1998) ou Clarke (1979), qui s'appuient sur la technique d'investigation des méprises (« miscues ») développée par Goodman (1969). D'après Bernhardt (1991, 22), la vision psycholinguistique de Goodman a régné en maître jusqu'en 1988, date à laquelle la publication de Carrell *et al.* (1989) diffuse véritablement les modèles interactifs au sein des chercheurs anglophones en lecture dans une L2 [29]. Cette prédominance des modèles descendants dans la recherche en L2 étonne d'autant plus que ces modèles n'ont jamais été dominants parmi les chercheurs en L1 et que ceux-ci les ont rapidement remis en question.

Du côté francophone, l'approche globale a également connu un vif succès sous l'impulsion de Moirand (1976, 1979) ou Lehmann et domina tout autant les années 70 à 90, à un point tel que Lehmann (1994, 90), parlant des travaux de Vigner (1979), Moirand (1979) et d'autres, a pu dire :

> *[ils] font en effet preuve (...) d'une extrême pudeur bibliographique ; comme si l'évidence était tellement aveuglante qu'ils serait inconvenant d'y insister autrement que par de rares et discrètes références aux seuls Smith et Goodman.*

Ce n'est que lentement que s'opère le glissement vers le paradigme qui domine alors les études en L1 : les premières références à Rumelhart (1977) se rencontrent déjà chez Cornaire (1985) et les modèles interactifs sont répandus par Cicurel (1991).

Dans les recherches en L1, comme nous l'avons dit, le paradigme « top-down », et en particulier le modèle de Goodman (1967), a par contre rapidement fait l'objet de critiques parfois vives [30], tandis que plusieurs de ses postulats étaient remis en question par des expérimentations postérieures. Ainsi, la notion de goulot d'étranglement a laissé place à une vision du lecteur expérimenté qui est capable de traiter rapidement un grand nombre d'informations visuelles (Rayner *et al.*, 2001, 45). Mais c'est surtout la conviction selon laquelle le bon

29. Bernhardt (1991, 22-25) propose les deux explications suivantes à ce phénomène : « It may be that academicians in this field have agreed that the psycholinguistic framework provides the most viable explanation of reading a second language. Another is that there is a basic lack of awareness and perception of the capabilities of models other than those of Goodman (1968) and Smith (1971) to explain second language reading phenomena ».

30. Ainsi, Kozloff (2002) démonte de manière systématique, non seulement les arguments de Goodman, mais aussi la scientificité de son raisonnement, affirmant : « He [Goodman] provides no logical, empirical, or commonsensical support for this conception. He then presents a highly selective set of passage misreadings by a child. These misreadings are not called errors ; they are "miscues". These misreadings are interpreted in a way that fits Goodman's guessing game formulation (although other interpretations — from the phonic and word centered approaches he disparages — are more obvious and reasonable) » Kozloff (2002, 35).

lecteur émet plus d'hypothèses et recourt davantage au contexte lors de la reconnaissance lexicale qui a été attaquée [31]. Tout d'abord, Stanovich (1980, 34) a mis en doute la possibilité que concevoir une hypothèse pour le mot suivant soit plus rapide que de simplement le reconnaître à partir d'informations visuelles. De plus, il rapporte une série d'études qui révèlent que, aussi bien au niveau orthographique qu'au niveau de la phrase, les bons lecteurs, au contraire des lecteurs plus faibles, utilisent en fait moins la redondance de la langue et moins le contexte pour reconnaître les mots. Eskey (1998, 95) renchérit :

> *They can decode, (...) both the lexical units and syntactic structures they encounter in texts, and they do so, for the most part, not by guessing from context or prior knowledge of the world, but by a kind of automatic identification that requires no conscious cognitive effort.*

Il conclut ainsi « the use of this top-down strategy is not, as top-down theorists would have us believe, the hallmark of good reading in every situation » (Eskey, 1998, 95).

Ces dernières remarques paraissent revenir vers une conception ascendante de la lecture, où la perception et les mécanismes de décodage jouent un grand rôle. Dans ce débat entre les deux paradigmes, on pourrait penser que l'approche ascendante sort vainqueur. Urquhart et Weir (1998, 44) déclarent en effet : « In fact it is virtually accepted in psychology nowadays that, at least at the level of word recognition and lexical access, some form of bottom-up process is followed ». Toutefois, les modèles descendants ont tout de même apporté plusieurs contributions essentielles à la recherche sur la lecture, en particulier la mise en évidence et la description des traitements qui s'opèrent au-delà du niveau du mot, un sujet sur lequel aussi bien Gough (1972) que LaBerge et Samuels (1974) restent très vague. Cet apport sera intégré à l'approche ascendante au sein des modèles interactifs.

Les modèles interactifs

Le premier à introduire ce nouveau paradigme est Rumelhart (1977). Il estime que le formalisme du traitement séquentiel de l'information utilisé par l'approche « bottom-up » ne permet pas de décrire adéquatement certaines données expérimentales. Parmi ces faits, il rapporte une perception plus rapide des lettres lorsqu'elles sont incluses dans un mot ou une suite de lettres respectant les règles orthographiques de l'anglais que lorsqu'elles sont isolées (c'est l'effet de supériorité du mot) (Huey, 1908; Reicher, 1969), la tendance des lec-

31. À ce sujet, il faut distinguer deux emplois distincts du terme contexte : « (...) the use of prior knowledge (top-down processing) to facilitate the simple recognition of words, and the use of such knowledge to facilitate higher-level interpretations of texts. The latter is characteristic of good readers, but the former is not » (Eskey, 1998, 95). C'est évidemment la première de ces deux significations qui constitue le coeur des modèles descendants.

teurs à commettre des erreurs de substitution où la catégorie de discours du mot remplaçant correspond à celle du mot remplacé (Kolers, 1970; Weber, 1970) ou encore l'effet d'amorçage qu'a un mot sur la vitesse de reconnaissance d'un mot qui lui est sémantique proche (voir notamment Meyer, Schvaneveldt and Ruddy, 1974, cités par Rumelhart 1977). Toutes ces observations sont interprétées comme des preuves de l'influence de processus de niveau supérieur sur la reconnaissance des mots. De même, Rumelhart (1977) produit d'autres faits mettant en évidence la dépendance de l'analyse syntaxique envers le contexte sémantique ou de l'interprétation sémantique envers le contexte général du texte.

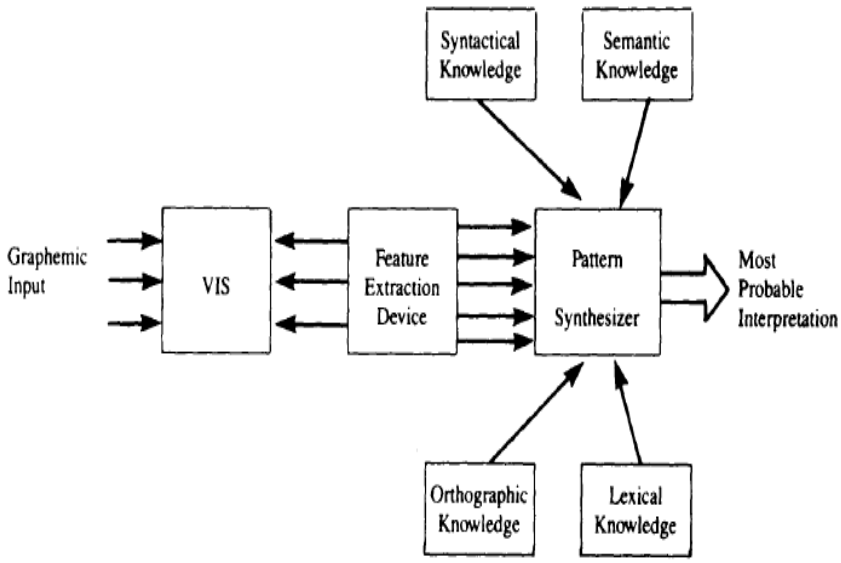


Figure 1.2 – Modèle de Rumelhart (1977)

Dès lors, afin de mieux interpréter ces résultats expérimentaux, Rumelhart (1977) propose un formalisme inspiré de la programmation parallèle (*cf.* Figure 1.2). Il conçoit un composant, le « Pattern Synthesizer », qui recueille à la fois les informations en provenance d'un dispositif d'extraction des traits — ayant analysé l'information graphémique préalablement enregistrée dans une réserve sensorielle, la VIS [32] — et de quatre niveaux de connaissances propres au lecteur : orthographique, lexicale, syntaxique et sémantique. Le modèle permet ainsi une interaction entre les composants des niveaux supérieurs et inférieurs : le sens se construit à partir de la perception visuelle, comme dans les modèles ascendants, tout en étant influencé par le contexte [33].

32. Le VIS (« Visual Information Store ») a été décrit par Sperling (1960, cité par Reicher 1969, 275) comme un système conservant un bref instant la perception visuelle et doté d'une capacité limitée, puisque les sujets ne sont généralement pas capables de rapporter l'ensemble des informations qui y sont emmagasinées avant leur déliquescence.

33. À l'inverse, ce qui sépare les modèles descendants des modèles interactifs est clairement explicité par Stanovich (1980, 35) : « in top-down models, semantic processes direct lower-level processes, whereas in in-

Toutefois, il est bien conscient des limites de son modèle, à propos duquel il déclare : « it is of very little help as a model of reading. (...) All that is interesting in the model takes place in the box labelled “Pattern Synthesizer” » (Rumelhart, 1977, 588-589). Il cherche bien à décrire la nature de la communication entre les cinq sources d’information à l’aide d’un mécanisme qu’il appelle « message center » (Rumelhart, 1977, 589) :

> *The message center keeps a running list of hypotheses about the nature of the input string. Each knowledge source constantly scans the message center for the appearance of hypotheses relevant to its own sphere of knowledge. Whenever such a hypothesis enters the message center the knowledge source in question evaluates the hypothesis in light of its own specialized knowledge. As a result of this analysis, the hypothesis may be confirmed, disconfirmed, and removed from the message center. This process continues until some decision can be reached.*

Ces développements ne convainquent pas Rayner et Pollatsek (1989, 467), pour lesquels le modèle de Rumelhart (1977) ne parvient pas à dresser un tableau de l’ensemble des processus de la lecture : il ne précise pas sur quelles bases sont générées les différentes hypothèses, ne spécifie pas l’importance relative des différentes sources d’informations et élude des problématiques telles que l’emploi de la voie phonologique pour la reconnaissance des mots ou la compréhension à un niveau supérieur que la phrase.

Par la suite, d’autres modèles interactifs sont apparus : Just et Carpenter (1980); Kintsch et van Dijk (1978); Samuels et Kamil (1984); Stanovich (1980). Ils considèrent tous plus ou moins que pour « identifier la signification des mots le lecteur pourrait faire appel à de nombreuses sources d’informations, aussi bien graphémiques, lexicales, morphologiques, syntaxiques, qu’à des connaissances plus générales comme par exemple des connaissances relatives au fonctionnement et à l’organisation des textes ou encore les connaissances portant sur les domaines référentiels des textes. » (Cornaire, 1999, 24). De plus, la majorité de ces modèles sont influencés par la théorie des schémas présentée précédemment.

Ainsi, Kintsch et van Dijk (1978) proposent le premier modèle des processus de compréhension, envisagés comme la construction d’une structure sémantique cohérente selon un processus cyclique contraint par les capacités de la mémoire de travail. Ce modèle, déjà introduit à la Section 1.2.2, est exposé en détail dans le Chapitre 2. Disons simplement ici qu’il appréhende un texte sous la forme d’une liste de propositions et que comprendre

teractive models semantic processes constrain the alternatives of lower levels but are themselves constrained by lower-level analyses ». Eskey (1998, 94) ajoute, parlant du point de vue du lecteur : « this so-called interactive model does not presuppose the primacy of top-down processing skills — the gradual replacing of painful word-by-word decoding with educated guessing based on minimal visual cues (...). In this view, good readers are both good decoders and good interpreters of texts, their decoding skills becoming more automatic but no less important as their reading skill develops ».

revient à les agencer en une structure cohérente, aussi bien au niveau microstructurel que macrostructurel. Suite aux limitations mémorielles des lecteurs, plusieurs cycles de traitement sont nécessaires pour mener cette tâche à bien. Durant chaque cycle, deux opérations cognitives sont principalement effectuées : des inférences, qui servent à associer deux propositions sans liens explicites et des réinstallations, qui visent à connecter une proposition isolée à du contenu en mémoire à long terme.

Stanovich (1980) étend en quelque sorte le modèle de Rumelhart (1977) en introduisant l'hypothèse compensatoire. Comme Rumelhart (1977), il reconnaît que le lecteur utilise simultanément des informations issues de sources diverses, mais il ajoute que le déficit d'un processus particulier est compensé par une plus grande dépendance vis-à-vis de connaissances issues d'autres niveaux de traitement. Ainsi, un lecteur confronté à une écriture peu lisible et qui ne peut s'appuyer totalement sur ses mécanismes de décodage, aurait la possibilité de faire appel à des éléments contextuels pour reconnaître les mots peu lisibles. Cette hypothèse permet ainsi d'expliquer la tendance, considérée jusqu'alors comme anormale au vu des théories descendantes, des lecteurs plus faibles à s'appuyer davantage sur le contexte pour effectuer des tâches de reconnaissance des mots (Allington et Fleming, 1978; Schvaneveldt *et al.*, 1977; West et Stanovich, 1978) [34], puisqu'il représente une source supplémentaire d'informations.

À la même époque, Just et Carpenter (1980) tentent de spécifier plus précisément les modèles interactifs en se servant de la durée des fixations visuelles mesurée au moyen d'un tachistoscope. S'appuyant sur des études qui révèlent la tendance du lecteur à fixer la majorité des mots, ils justifient la validité de leur protocole expérimental à l'aide des deux présupposés suivants : d'une part, l'interprétation de ce qui est perçu est immédiate (« immediacy assumption ») ; d'autre part, les yeux restent fixés sur un mot aussi longtemps que celui-ci est traité (« eye-mind assumption »). Par conséquent, le temps de fixation devrait être un indice fiable du temps de traitement cognitif.

Sur la base de cette méthode d'investigation, les auteurs développent un modèle de régression linéaire multiple qui permet de prédire la durée de fixation d'un mot en fonction d'une série de variables appartenant à différents niveaux de traitements : lexical, syntaxique et sémantique (Just et Carpenter, 1980, 337) et donc de décrire en temps réel les processus

34. À ce sujet, Stanovich (1980, 42) distingue deux types de processus liés au contexte. Le premier concerne l'élaboration d'une structure de connaissances à partir du texte — ce qui correspond en gros à la compréhension du texte — et s'est révélé employé plus efficacement par des lecteurs expérimentés (Guthrie, 1973; Smiley *et al.*, 1977). Le second correspond à ce que les modèles descendants désignent avant tout sous le terme d'hypothèse contextuelle de lecture (Goodman, 1967; Levin et Kaplan, 1970; Smith, 1971), à savoir le support contextuel pour la reconnaissance des mots. C'est cette seconde conception du contexte qui est ciblée par la remarque de Rumelhart et à propos de laquelle les diverses études citées dans le texte principal ont montré que les lecteurs plus faibles ont davantage tendance à utiliser le contexte lors d'une tâche de prononciation de mots.

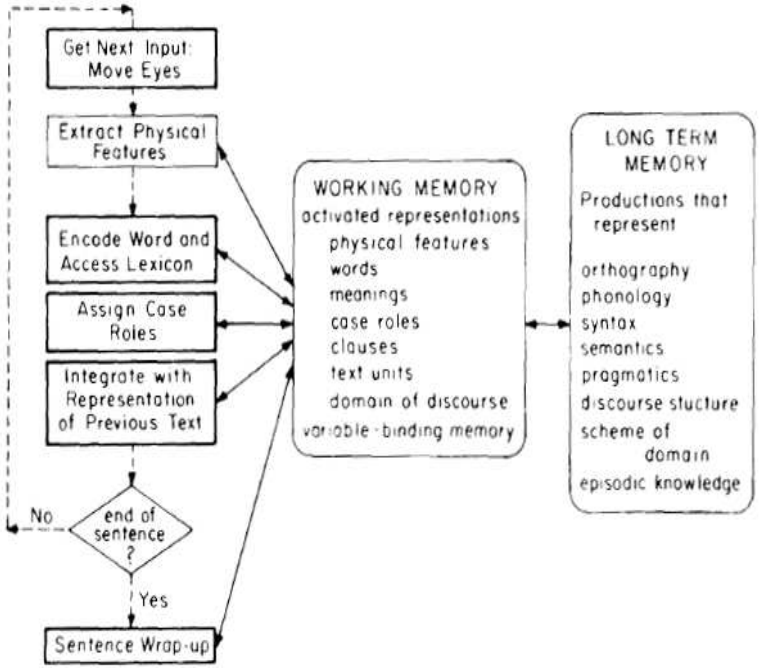


Figure 1.3 – Modèle de Just and Carpenter (1980)

de lecture. Toutefois, l'aspect le plus important de leurs travaux tient sans doute à l'architecture de leur modèle (*cf.* Figure 1.3), composée de différents niveaux de traitements, de la mémoire de travail (MT) et de la mémoire à long terme (MLT). C'est dans la MT, où parviennent les informations extraites du texte, que se construit l'interprétation, sous la guidance d'une série de règles de productions récupérées dans la MLT [35]. Ces règles encodent les propriétés orthographiques, phonologiques, syntaxiques, sémantiques de la langue ainsi que des schémas propres à un thème ou un type de discours précis. Quel que soit le niveau de traitement auquel elles appartiennent, elles produisent une action aussitôt que les concepts contenus dans la MT remplissent certaines conditions, ce qui rend l'architecture interactive et conforme à l'hypothèse d'immédiateté [36]. De plus, cet ordonnancement autorise un niveau de formalisation plus élevé du modèle : Just et Carpenter (1980) concluent d'ailleurs en envisageant des simulations informatisées des processus de la lecture.

Au niveau des recherches en L2, comme nous l'avons dit plus haut, il a fallu attendre longtemps avant que les modèles interactifs tels que nous venons de les présenter, soient adoptés par cette communauté, alors que le terme interactif a très vite prospéré dans un second sens. Certains auteurs l'ont en effet utilisé pour désigner l'interaction entre les connaissances que contient le texte et celles du lecteur, une perspective qui provient à la fois du modèle psy-

35. Remarquons que dans ce modèle, la MT remplit un rôle fort similaire à celui du « message center » de Rumelhart (1977). L'architecture de Just et Carpenter (1980) respecte toutefois davantage les connaissances de l'époque en matière mémorielle (Atkinson et Shiffrin, 1968; Baddeley, 2003).

36. Celle-ci a toutefois été remise en question par des expérimentations postérieures qui tendent à conclure que les traitements de haut niveau débordent sur la fixation suivante (Rayner et Pollatsek, 1989, 471).

cholinguistique de la lecture de Goodman que des travaux sur les schémas (Grabe, 1998). Cette approche a produit une série d'études sur les influences diverses de connaissances telles qu'un bagage à propos de la thématique du texte (Carrell, 1983), le fait de savoir de quoi parle le texte (Carrell, 1983), les connaissances culturelles des apprenants (Steffensen *et al.*, 1979), voire la manière dont l'apprenant envisage les mécanismes de lecture (Devine, 1984).

Les modèles interactifs, aussi bien en L1 qu'en L2, constituent la seule organisation du système de lecture qui soit soutenable de nos jours (Ferrand, 2007, 19). Elle connaît toutefois des variations. La structure en cascade de McClelland (1979) n'est pas à proprement interactive (il n'y a pas de feedback), mais autorise différents sous-modules à travailler en même temps en postulant un flux continu d'informations qui passe de l'un à l'autre. L'organisation la plus commune (McClelland et Rumelhart, 1981) considère toutefois l'existence d'interactions entre les différents processus impliqués lors de la lecture, que ce soit par cascade, par feedback ou au moyen de connexions latérales (c'est-à-dire entre des composants d'un même niveau de traitement). Elle conduira à des modèles informatisés tels que le paradigme par traitement parallèle distribué (en anglais Parallel Distributed Processing (PDP)) (Rumelhart *et al.*, 1986b; Seidenberg et McClelland, 1989), dont nous parlerons par la suite, ou les modèles à deux routes (Coltheart *et al.*, 1993, 2001).

Les modèles connexionistes

Les modèles connexionnistes sont basés sur l'hypothèse que la complexité peut naître de l'agrégation de larges réseaux composés d'unités de traitement simples (Seidenberg, 2006, 235) et sur une correspondance avec les structures cérébrales humaines[37]. Celles-ci sont décrites brièvement par Tan *et al.* (2005, 246) :

> *The humain brain consists primarily of nerve cells called neurons, linked together with other neurons via strands of fiber called axons. Axons are used to transmit nerve impulses from one neuron to another whenever the neurons are stimulated. A neuron is connected to the axons of other neurons via dendrites, which are extensions from the cell body of the neuron. The contact point between a dendrite and an axon is called a synapse.*

Ainsi, l'approche connexionniste a transposé les circuits neuronaux sous la forme de nœuds (les neurones) reliés par des connexions orientées et pondérées (représentant les axones et les dendrites) et organisés en couches, dont le niveau inférieur représente l'input et le niveau

37. La prémisse à la base de cette équivalence est « qu'une structure semblable au cerveau devrait avoir des capacités semblables à celui-ci » (O'Reilly and Munakata, cités par Fiset et Arguin 2002, 189).

supérieur, l'output.

L'analogie ne s'arrête pas là : l'activation et l'apprentissage de ce modèle statistique reposent également sur deux principes biologiques. D'une part, McCulloch and Pitts (1943, 1947, cités par Larson 2004) ont interprété le principe d'activation des neurones comme étant la somme des informations en provenance d'autres neurones. Par conséquent, dans ce formalisme, le niveau d'activation d'une unité de traitement sera calculé comme la somme des activations pondérées des neurones du niveau inférieur qui lui sont connectés. D'autre part, Hebb (1949, cité par Larson 2004) a expliqué comment l'apprentissage est le résultat de modifications au niveau des connexions synaptiques : en effet, plus la synapse transmet d'informations entre deux neurones et plus leur connexion tend à se renforcer physiquement. À l'inverse, si peu d'informations transitent via cette synapse particulière, la connexion va s'affaiblir, voire disparaître. Cette particularité biologique a été transposée au moyen de poids, associés à chaque connexion, qui expriment l'aptitude variable qu'ont les informations issues d'un nœud à affecter un nœud voisin (Page, 2000, 444).

On considère généralement que la première application des réseaux de neurones à la lecture est due à McClelland et Rumelhart (1981), même si les auteurs s'inspirent fortement du modèle logogen de Morton (1969, cité par Randall 2007, 57) [38]. Il s'agit d'une version assez simple qui vise à expliquer comment le contexte peut influencer la perception (c'est-à-dire, dans ce cas-ci, qu'on reconnaît mieux une lettre dans le contexte d'un mot que lorsqu'elle est seule). Le formalisme mathématique y est encore relativement peu développé : le réseau ne comprend pas de couches de neurones cachées [39] et les poids attribués aux connexions, ainsi que divers autres paramètres, ont été définis ad hoc. De plus, bien que le modèle théorique complet (McClelland et Rumelhart, 1981, 377-378) envisage cinq niveaux de traitement, un accès au lexique aussi bien via la voie directe que phonologique et une influence de connaissances issues de niveaux supérieurs, seuls trois niveaux de connaissances ont été effectivement implémentés : les traits constitutifs des lettres, les lettres et les mots (*cf.* Figure 1.4). Le modèle se limite donc à simuler la reconnaissance des mots par la voie lexicale, sans prendre en compte ni la voie phonologique ni les processus syntaxiques ou sémantiques. Enfin, pour des raisons de taille du dictionnaire et du nombre de connexions, seuls les mots de quatre lettres sont traités.

38. Dans ce modèle, la reconnaissance d'un mot s'effectue au moyen d'un détecteur, un « logogen », spécifique à ce mot, lequel est activé par diverses informations fournies au système. Passé un certain seuil d'activation, le mot est reconnu. De plus, ce modèle explique élégamment l'effet de fréquence par une variation du niveau d'activation au repos des logogens. Ainsi, un logogen dédié à un mot fréquent nécessite moins d'informations avant de passer son seuil d'activation. On retrouve dans ce formalisme plusieurs des concepts qui feront le succès de l'approche connexionnistes.

39. Or, comme l'a fait remarquer Hinton (1989, 193), l'absence de couches cachées entraîne des limitations dans les capacités de traitement des réseaux de neurones, notamment parce qu'ils sont alors incapables de modéliser des frontières non linéaires entre différentes classes.

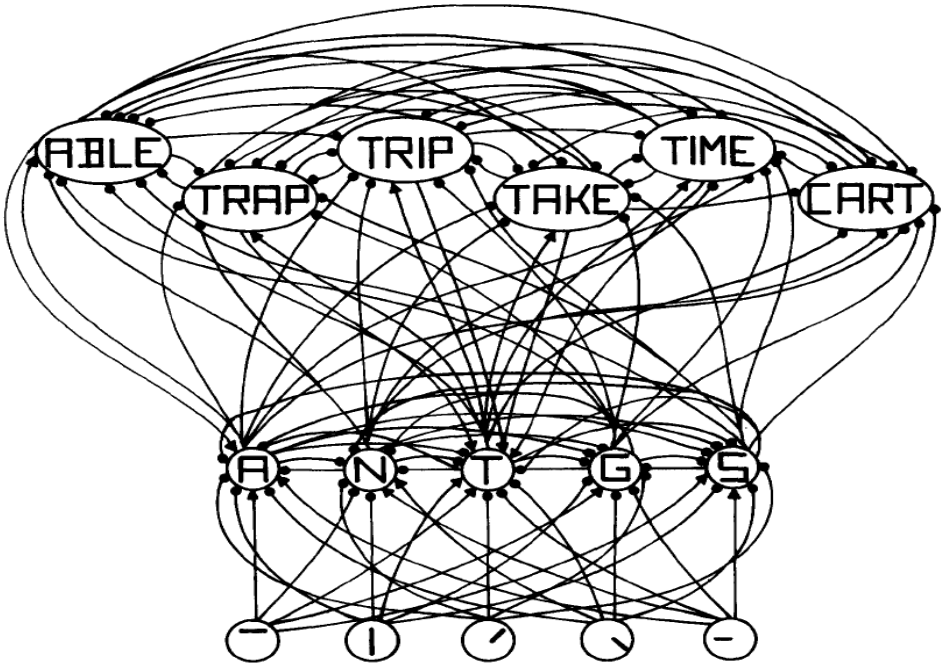


Figure 1.4 – Modèle de McClelland and Rumelhart (1981) : les voisins de la lettre T.

À la décharge des auteurs, il faut reconnaître que la tâche n'est pas simple. En effet, cette modélisation se nourrit des principes interactifs : la transmission de l'information s'effectue en cascade, le traitement des données se fait en parallèle à différents niveaux et les connections entre nœuds sont bidirectionnelles, permettant aussi bien aux informations de niveau supérieur que de niveau inférieur d'influencer l'activation des neurones (Plaut, 2005, 30). Dès lors, le nombre de connexions entre tous les nœuds du modèle est important et justifie les différentes simplifications pour lesquelles ont opté McClelland et Rumelhart (1981). Par ailleurs, les auteurs reprennent à Levin (1976, cité par McClelland et Rumelhart 1981) une notion essentielle pour les réseaux connexionnistes : la capacité d'un nœud à exciter ou à inhiber ses voisins, ce qui permet à deux nœuds susceptibles de coexister (par exemple, la lettre *l* et le mot *le*) de se renforcer l'un l'autre ou, inversement, à deux hypothèses inconciliables (par exemple, la lettre *t* et le mot *le* ou les mots *le* et *du*) de s'exclure.

Par la suite, différentes améliorations de ce formalisme vont être proposées au sein du paradigme par traitement parallèle distribué (Rumelhart *et al.*, 1986b). Tout d'abord, s'ouvre un débat autour de la question du type de représentations qu'il convient d'employer. Chez McClelland et Rumelhart (1981), la représentation des connaissances est locale, c'est-à-dire que chaque unité correspond à une entité familière, telle qu'une lettre, un phonème, un mot ou un concept. Par contre, les modèles PDP adoptent plutôt des représentations distribuées, c'est-à-dire où chaque entité est représentée par un motif d'activation particulier réparti entre un grand nombre de nœuds et non associé à un nœud particulier (Plaut, 2005, 27), c'est-à-dire, plus concrètement, par un ensemble de traits. Cette nouvelle conception de l'agencement

du cerveau humain rompt avec l'idée que certaines cellules, désignées dans la littérature sous l'expression « cellules grand-mère », puissent se spécialiser dans la reconnaissance de certains objets ou concepts (Page, 2000, 445). Toutefois, Page (2000) souligne que ce nouveau type de représentations n'est pas si radicalement différent des représentations locales, puisqu'une entité qui est représentée localement à un niveau n peut être considérée comme représentée de façon distribuée au niveau $n-1$. Par conséquent, ce sont davantage les niveaux supérieurs qui déterminent le type de représentations. Pour sa part, Seidenberg (2006, 236) soutient qu'une représentation n'est dite locale ou distribuée qu'en référence à d'autres unités et que donc, ce qui importe est de distinguer les modèles qui représentent les mots de ceux qui ne disposent pas de représentations à ce niveau.

Par ailleurs, différents travaux (dont Minsky and Papert, 1969, cités par Rumelhart *et al.* 1986a, 319) ayant montré les limitations des réseaux de neurones à deux couches — l'une pour les inputs, l'autre pour les outputs —, le paradigme PDP prône dès lors l'emploi de couches cachées entre les niveaux d'entrée et de sortie des données (Rumelhart *et al.*, 1986a). Cette architecture plus compliquée rend les modèles connexionistes capables de traiter des correspondances plus complexes et moins systématiques entre deux niveaux de connaissances. Toutefois, un problème important subsiste. Comment déterminer les poids de chacune des connexions, étant donné que les unités abstraites de la couche cachée se prêtent mal à un réglage manuel, qui, de toute façon risque d'être fort peu optimal ? Le développement d'un algorithme de type supervisé qui soit capable d'assigner aux poids des valeurs selon un optimum local constitue le troisième apport majeur du paradigme PDP à la psychologie cognitive et même plus généralement à la modélisation statistique (Rumelhart *et al.*, 1986a). On le connaît sous le nom d'algorithme de rétropropagation et vise à minimiser l'erreur entre les sorties du réseau et la réponse correcte en adaptant les poids du réseau suite à la rétropropagation de cette erreur vers les couches précédentes.

Le paradigme PDP n'a aujourd'hui rien perdu de son importance. Il a influencé nombre de modèles postérieurs, constituant autant d'améliorations et d'extensions du modèle de McClelland et Rumelhart (1981). Ainsi, Seidenberg et McClelland (1989) détaillent une architecture triangulaire, assez proche de celle de McClelland et Rumelhart (1981), qui comporte trois composantes, interconnectées via des couches cachées, et représentant les connaissances orthographiques, phonologiques et sémantiques (*cf.* Figure 1.5). Toutefois, seules les correspondances entre les formes graphiques et phonologiques ont été implémentées (par conséquent, il s'agit plutôt d'un modèle de lecture orale que de lecture silencieuse). Pour ce faire, ils recourent à des représentations distribuées. Leur problème principal est de traiter, au sein d'un formalisme qui n'est pas temporel par nature, la séquentialité de la langue (il faut que des mots comme *rien* et *nier* activent de manière différente les unités de la couche des représentations orthographiques). McClelland et Rumelhart (1981) avaient résolu cette

difficulté en se limitant à des mots de quatre lettres et en spécifiant un jeu complet d'unités (traits graphiques ou lettres) pour chacune des positions possible dans un mot. Seidenberg et McClelland (1989, 528) ont tenté une autre approche : ils ne tiennent pas compte de la séquentialité des différents éléments constitutifs de l'input, mais utilisent des trigrammes de lettres pour les unités orthographiques, et des triplets de traits phonétiques pour la composante phonologique.

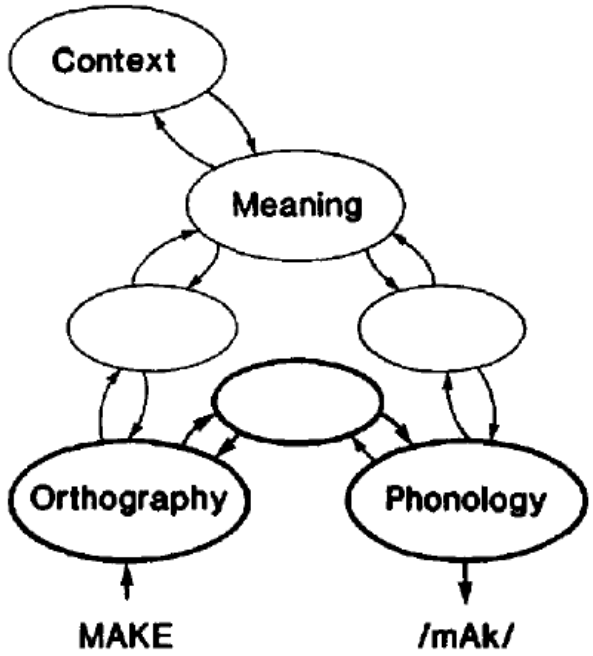


Figure 1.5 – Modèle de Seidenberg and McClelland (1989). Seules les parties en gras ont été implémentées.

Le modèle de Seidenberg et McClelland (1989) est ensuite complété par Harm et Seidenberg (2004), qui en implémentent la partie sémantique. Une fois achevé, il peut dès lors simuler les deux voies conventionnelles vers le sens : la voie directe (de l'orthographe à la sémantique) et la voie phonologique (de l'orthographe au sens via une représentation phonologique). Toutefois, à la différence des théories à deux voies décrites ci-dessous (Coltheart *et al.*, 1993, 2001), Harm et Seidenberg (2004) refusent de considérer ces deux chemins comme indépendants l'un de l'autre et en compétition. Au contraire, l'activation des unités sémantiques est à la fois le résultat des informations issues de la voie directe et de celles en provenance de la voie phonologique.

Ces différents modèles ont réussi, avec un succès variable, à reproduire une assez large gamme d'effets empiriques révélés par des études précédentes, acquérant de ce fait une solide réputation parmi les chercheurs en lecture. De plus, leur capacité de généralisation, c'est-à-dire l'habilité à traiter des formes jamais rencontrées, a été mise en évidence (Plaut,

2005, 25) : une fois entraînés, ils peuvent traiter des formes jamais rencontrées. Ils se sont enfin révélés utiles pour simuler les processus en cas de dommages cérébraux.

Diverses critiques ont pourtant été émises envers les modèles connexionistes. La plus évidente d'entres-elles est leur incapacité à prendre en compte les processus situés au-delà de la reconnaissance des mots. Plaut (2005, 30) explique que ce choix découle de l'adoption, par la majorité des chercheurs, de la conception selon laquelle, au-delà du niveau lexical, la compréhension du langage écrit fait globalement appel aux mêmes mécanismes que le langage oral. Une seconde critique, exprimée par Barber et Kutas (2007, 100) concerne l'incapacité des implémentations actuelles à traiter les mots plurisyllabiques. Troisièmement, les auteurs ajoutent que les réseaux de neurones et l'algorithme d'entraînement par rétropropagation n'est pas biologiquement plausible (2007, 101) : « it is thus commonly assumed that the brain uses less powerful learning rules in combination with more neurons as well as more complex patterns of connectivity ». De plus, l'architecture du cerveau humain ne se limite pas à un ensemble de neurones interconnectés. Il est notoire que le cerveau est divisé en deux hémisphères communiquant entre eux. Cette voie a toutefois été explorée, notamment par Weems and Reggia (cités par Barber et Kutas 2007, 104) qui ont développé un modèle neuronal comprenant deux couches cachées représentant chacun des hémisphères et capables de s'échanger de l'information.

Au niveau des recherches en L2, ce type de modèle ne semble pas avoir connu le même succès que les approches descendantes et interactives. Ainsi, dans sa section consacrée aux modèles connexionnistes, Koda (2005, 16-18), une spécialiste de la lecture en L2, ne cite guère que l'article de Bereiter (1991), qui vulgarise le paradigme connexioniste, sans faire référence aux articles majeurs de ce paradigme. Elle énumère toutefois quelques implications importantes des réseaux neuronaux pour l'étude de la lecture en L2. D'une part, l'approche connexionniste considère l'acquisition de compétences langagières dans une L2 comme le résultat de l'ajustement et du renforcement des connexions entre neurones suite à une série d'inputs. Cette conceptualisation se révèle très utile pour expliquer les effets que peut engendrer la distance entre la L1 et la L2 dans le processus de lecture. Ainsi, dans le cas de deux langues proches, l'adaptation des circuits neuronaux requise pour une pratique efficace de la L2 est légère et ne nécessite qu'une quantité limitée d'expériences (autrement dit d'input), alors que, dans le cas les deux langues éloignées, il faudra que l'apprenant soit longuement exposé à la langue cible avant d'adapter à celle-ci ses structures cognitives. D'autre part, les modèles connexionnistes expliquent l'automaticité comme la récupération indélibérée en mémoire d'un motif d'activation particulier (Koda, 2005, 18). Dès lors, il est hautement plausible que des circuits propres à la L1 d'un lecteur, régulièrement utilisés, soient activés par un input de la L2, certaines études ayant démontré cette interférence au niveau lexical (Van Heuven *et al.*, 1998).

Van Heuven *et al.* (1998) ont par ailleurs développé l'un des rares modèles connexionnistes décrivant la reconnaissance des mots dans un cadre bilingue : le modèle BIA. Celui-ci s'attaque à la question suivante : sachant qu'un lecteur bilingue possède, soit deux lexiques indépendants, soit un lexique intégré, que se passe-t-il lorsqu'il lit un mot dans l'une de ses deux langues ? Existe-t-il un mécanisme qui joue le rôle d'un interrupteur, permettant le transfert des informations visuelles, tantôt vers l'un des systèmes de reconnaissance lexicale, tantôt vers l'autre ? Le stimulus n'active-t-il pas plutôt des mots appartenant aux deux langues à la fois, étant donné que des langues proches comprennent souvent un certain nombre d'items orthographiquement similaires tels les homographes interlexicaux ou les congénères [40] ?

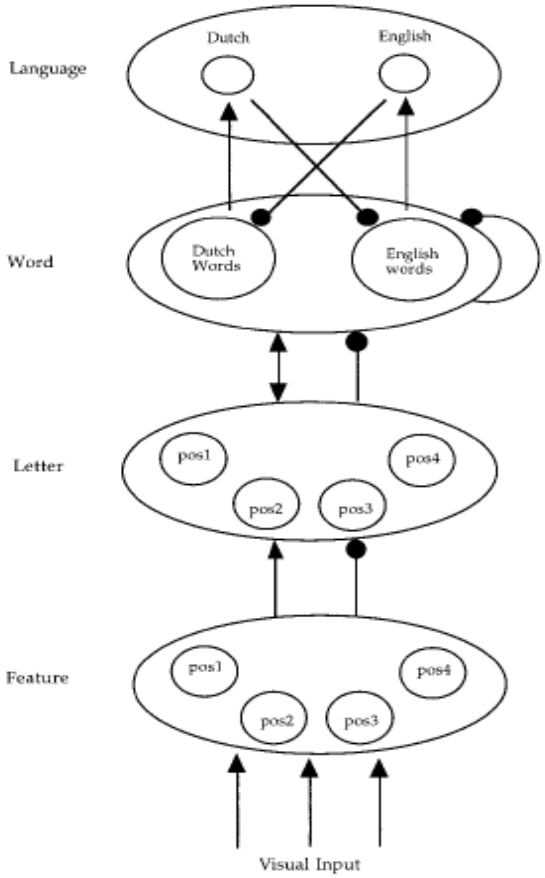


Figure 1.6 – Le modèle BIA (Bilingual Interactive Activation) de van Heuven, Dijkstra et Grainger (1998).

Les résultats obtenus par Van Heuven *et al.* (1998) tendant à démontrer la vraisemblance de la seconde possibilité : le bilingue posséderait un lexique intégré dont l'accès se ferait de manière non sélective. Dès lors, pour concevoir leur modèle de la reconnaissance des

40. Van Heuven (2005, 262) définit les congénères comme des mots orthographiés de manière similaire dans deux langues et de sens semblable, tandis que les homographes interlexicaux, ou faux-amis, sont des mots orthographiés de manière similaire, mais présentant un sens différent (ex. sol (esp.) et sol (fr.).

mots chez les bilingues (*cf.* Figure 1.6), les auteurs reprennent l'architecture du modèle IA de McClelland et Rumelhart (1981), à laquelle ils ajoutent deux extensions : d'une part, un lexique intégré comprenant des mots de l'anglais et du néerlandais ; d'autre part, une couche supplémentaire contenant deux nœuds langagiers, connectés avec l'ensemble des éléments du lexique, et dont la fonction principale est de collecter l'activitation en provenance de lexèmes de la langue qu'ils représentent tout en inhibant les mots de l'autre langue (Van Heuven *et al.*, 1998, 476). Ils parviennent dès lors à expliquer un ensemble de phénomènes d'interférences observés dans diverses expériences. Le modèle BIA, s'appuyant sur le modèle IA, partage toutefois les limitations de ce dernier (Van Heuven, 2005, 274). Ainsi, il ne comporte pas d'informations phonologiques ou sémantiques, ce qui limite considérablement son pouvoir explicatif et n'accepte que des mots de 4 lettres comme input. Enfin, le principe d'inhibition des mots par feedback d'un nœud langagier ne semble pas correspondre à un mécanisme existant. C'est pourquoi il a connu une seconde version, BIA+ (Dijkstra and Van Heuven, 2002, cités par Van Heuven 2005), que nous ne développerons pas ici. L'essentiel de notre propos étant ici de montrer un exemple de l'application du paradigme connexionniste pour rendre compte des variations dues à la langue maternelle des apprenants.

Le modèle à deux voies

Le modèle à deux voies se caractérise avant tout par la conviction que le lecteur expérimenté dispose de deux voies pour convertir l'écrit vers l'oral : la voie lexicale, qui accède à la prononciation via un lexique mental, et la voie phonologique, qui opère par une série de règles de correspondances entre les lettres et les sons (Coltheart *et al.*, 1993, 589). La première voie, présumée la plus utilisée, permet de lire la majorité des mots pourvu qu'ils soient déjà connus du lecteur ; la seconde intervient pour les mots nouveaux ou les non-mots et est susceptible d'engendrer des erreurs dans le cas de vocables à la prononciation irrégulière. Cette hypothèse opère une dichotomie forte entre les deux voies : « seule une route permet l'accès au lexique, en règle générale, la route la plus rapide » (Ferrand, 2007, 199), c'est-à-dire, dans la majorité des cas, la voie lexicale. Il s'agit d'une différence notoire par rapport à la conception de Harm et Seidenberg (2004), lesquels postulent la coopération des deux routes de traitement, toutes deux susceptibles d'activer la composante sémantique du modèle.

Cette conception duelle de la lecture repose sur une longue tradition : le premier à l'envisager pour la lecture est Saussure (1922, cité par Coltheart 2005, 6). Dans les années 70, elle se répand, en particulier sous l'effet de l'article de Marshall et Newcombe (1973), qui la décrivent selon une architecture modulaire, c'est-à-dire, sous la forme d'un diagramme en forme de boîtes et de flèches. Par la suite, les modèles à deux voies ont conservé cette archi-

tecture modulaire, à la différence des réseaux de neurones, qui privilégient des représentations distribuées. Fiset et Arguin (2002, 183) expliquent : « Les études neuropsychologiques menées auprès de patients cérébrolésés ont fortement contribué à établir la crédibilité de la théorie de lecture proposée par l'approche modulaire, qui implique deux voies de traitement distinctes. En fait, le modèle à deux voies a été élaboré en parallèle aux découvertes majeures faites par la neuropsychologie cognitive au sujet des dyslexies acquises ». Celles-ci constituent d'ailleurs déjà le sujet de l'article de Marshall et Newcombe (1973).

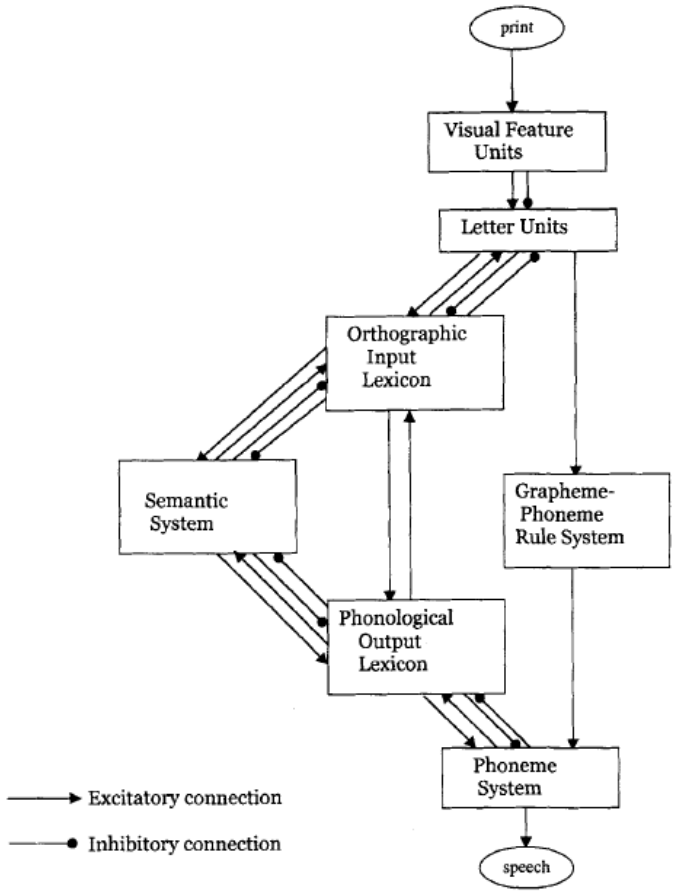


Figure 1.7 – Modèle DRC de Coltheart, Rastle, Perry, Langdon and Ziegler (2001).

Parmi les divers modèles à deux voies, le plus connu est le Dual Route Cascade (DRC), qui s'appuie sur les modèles logogen de Morton (1969, 1979, 1982, cité par Ferrand 2007, 179) et sur le modèle **IAC!** (**IAC!**) de McClelland et Rumelhart (1981). D'abord centré sur la lecture à haute voix (Coltheart *et al.*, 1993), il a ensuite été étendu à la lecture silencieuse (Coltheart *et al.*, 2001). Le modèle DRC (*cf.* Figure 1.7) conçoit la lecture comme une série de processus, organisés selon une architecture à deux routes [41] spécifiée manuellement, qui

41. Il propose en fait trois routes, étant donné que la voie lexicale se présente sous deux formes : la chaîne orthographique peut en effet activer une suite phonologique, soit directement, soit via le passage par un lexique mental qui contient, notamment, des informations phonologiques.

vont du traitement des traits visuels à la production des sons. Chaque niveau encode une classe d'informations symboliques par un ensemble d'unités : il s'agit de représentations locales telles des traits, des lettres, des mots ou encore des phonèmes... Ces représentations sont reliées entre elles, tantôt par un ensemble de correspondances qui s'appliquent de manière sérielle (de gauche à droite) dans le cas de la route phonologique, tantôt, dans le cas de la route lexicale, par des connexions inhibitrices ou excitatrices dont l'activation se propage en cascade (cf. McClelland, 1979) et de manière interactive (mécanismes de feedback), et dont les poids ont été assignés manuellement. Le modèle DRC reproduit donc certaines caractéristiques des réseaux neuronaux, mais s'en distingue en ce qu'il ne recherche ni une plausibilité biologique (Coltheart, 2005, 16), ni une détermination automatique de sa structure.

Le modèle DRC a obtenu d'excellents résultats, en ce qu'il est capable de reproduire une large gamme d'effets observés chez l'humain (cf. Coltheart *et al.*, 2001). Il se comporte d'ailleurs mieux que les modèles connexionnistes quant à la prononciation des non-mots (Coltheart, 2005, 18), ce qui ferait de lui le meilleur modèle à l'heure actuelle. Toutefois, il présente lui aussi plusieurs limitations. Les succès obtenus dans l'explication de divers phénomènes ne suffissent pas à garantir que le fonctionnement et l'architecture du modèle correspondent bien au cerveau humain (Coltheart *et al.*, 2001, 204; Barber et Kutas, 2007, 103) . De plus, s'il est adapté aux langues alphabétiques, telles l'anglais, le français ou l'allemand, il ne l'est probablement pas pour des langues logographiques telles le japonais ou le chinois (Ferrand, 2007, 286). Par ailleurs, comme les modèles connexionnistes, il s'avère incapable de traiter les mots polysyllabiques. Notons enfin que DRC a été critiqué par Perry *et al.* (2007) à propos de certains phénomènes qu'il ne parvient pas à expliquer de manière satisfaisante. Les auteurs ont d'ailleurs proposé une version plus complexe, le modèle connexionniste à deux processus (CDP+), capable de simuler davantage d'effets.

Au niveau de la lecture en L2, la pertinence de DRC pour l'ensemble des langues étant déjà plus que sujette à caution, l'implémentation d'une version bilingue semble hautement complexe et n'existe pas actuellement, à notre connaissance. Par contre, des travaux s'appuient sur cette perspective théorique largement acceptée pour l'anglais dans la description qu'ils font des particularismes de la reconnaissance lexicale lors d'une lecture en L2 (Randall, 2007, 70-73). Notons aussi que Norris (1994, cité par Randall 2007, 74) a souligné le fait que les deux voies de DRC peuvent être utilisées à des degrés très divers selon la nature de la L1 et la L2 de l'apprenant. Reste aussi à déterminer si ce modèle correspond à une réalité neurologique, aussi bien dans le cas de la lecture en L1 qu'en L2. Pour répondre à cette interrogation, il nous faut maintenant nous tourner vers les observations réalisées à l'aide des dernières technologies en imagerie mentale.

L'apport des neurosciences

En intitulant cette partie *apport* des neurosciences, nous avons voulu insister sur le fait que cette section ne présente pas un paradigme théorique qui inspire de nouveaux modèles de lecture, mais plutôt une nouvelle voie d'investigation des processus cognitifs de la lecture, à l'instar des techniques de la chronométrie mentale ou des approches computationnelles. Les neurosciences laissent présager d'un grand pas en avant : en effet, il devient possible de comparer les constructions théoriques décrites précédemment avec la structure réelle du cerveau humain, afin d'obtenir des modèles qui soient biologiquement plausibles. L'importance que cette nouvelle voie d'investigation a pris récemment est telle qu'il apparaît justifié de lui consacrer une section.

En réalité, lorsqu'on traite des apports neurosciences, il convient de distinguer en leur sein deux approches différentes (Ferrand, 2007, 44-45), toutes deux trouvant leurs origines dans les travaux en neuropsychologie de la fin du XIX^{e} siècle (Broca, Déjérine, Wernicke...). La première d'entres-elles, la neuropsychologie cognitive, cherche à « mettre en relation les observations des patients cérébro-lésés avec les modèles du fonctionnement de l'être humain normal » (Ferrand, 2007). C'est ainsi que le modèle à deux routes s'est vu conforté par l'étude de patients aux cerveaux lésés.

La deuxième repose sur une série de techniques d'imagerie cérébrale et permet d'observer le cerveau en train de fonctionner. C'est sur cette dernière que nous nous concentrerons ici. Nous commencerons par résumer le fonctionnement de ces diverses techniques, avant de préciser quelles aires du cerveau jouent un rôle lors de la lecture, et quel est ce rôle.

La plus ancienne de ces techniques est la Tomographie par Émission de Positions (TEP) : elle a été utilisée pour l'étude des processus de lecture pour la première fois par Petersen *et al.* (1988) et relève du domaine de la physique nucléaire. Elle s'appuie sur le fait que les cellules du corps humain consomment d'autant plus de glucose qu'elles sont plus actives. Dès lors, pour réaliser une TEP, on incorpore un isotope radioactif à du glucose qui est ensuite absorbé par le corps. Cet isotope, entraîné par le glucose dans les régions actives de l'organisme, va rapidement s'y désintégrer en émettant un positon, c'est-à-dire un électron positif qui fait partie du domaine de l'antimatière et qui va se détruire dès qu'il rencontre un électron environnant. En se détruisant, ces deux particules se transforment en énergie sous la forme de deux photons γ, dont l'émission s'effectue dans des sens opposés. Reste alors aux détecteurs placés autour de la tête du patient à reconstituer la distribution de la radioactivité dans l'espace sur la base de ces photons γ et à cartographier ainsi les zones actives du cerveau à ce moment précis (Dehaene, 2007, 102).

Une technique plus récente, l'Imagerie Fonctionnelle par Résonance Magnétique (IRMf),

permet également de cartographier les zones actives du cerveau, mais sans risque pour la personne (car on n'injecte pas de produit radioactif cette fois). Elle requiert l'usage d'un large aimant en forme de tube, qui produit un champ magnétique permettant de mesurer les différences dans l'état d'oxygénation de l'hémoglobine au travers de l'effet Blood Oxygen Level Dependant (BOLD) (Goswami, 2004, 5). En effet, une molécule d'hémoglobine non oxygénée perturbe le champ magnétique, tandis qu'une molécule oxygénée entraîne une augmentation mesurable du champ magnétique (Dehaene, 2007, 105). Or, plus une zone du cerveau est active, et plus elle est irriguée par un afflux de sang oxygéné et donc, plus elle est détectable à l'IRMf. Par rapport à la TEP, cette technique présente une résolution temporelle bien plus élevée : une image toutes les deux ou trois secondes contre une image toutes les 10 ou 15 minutes (Dehaene, 2007, 106). Toutefois, toute deux conservent un certain temps de latence qui engendre une imprécision à propos du déroulement temporel de l'activité cérébrale.

Dès lors, pour mieux mesurer cet aspect temporel des processus cognitifs de la lecture, il existe deux autres techniques, plus anciennes : les Potentiels Électriques Évoqués (PEV) et la Magnéto-Encéphalographie (MEG). La technique des PEV « consiste à enregistrer les [différences de] potentiels électriques générés par les neurones en plusieurs points de la surface du crâne grâce à des électrodes (entre 64 et 256) fixées sur le cuir chevelu et reliées à un potentiomètre sensible » (Ferrand, 2007, 45). Elle présente une résolution temporelle élevée, mais une résolution spatiale faible, car les tissus traversés par le courant affectent la précision des mesures (Salmelin et Kujala, 2006, 520). La MEG mesure quant à elle « les minuscules variations de champ magnétique induites par ces mêmes courants » (Dehaene, 2007, 114-115). Plus chère à mettre en œuvre, elle reflète aussi plus fidèlement l'activité du cerveau, puisque « les signaux magnétiques sont moins distordus par les tissus traversés que les signaux électriques » des PEV (Ferrand, 2007, 44).

À l'aide de ces quatre méthodes d'investigation, les experts en neurosciences ont cherché à établir à la fois la cartographie des zones actives lors de la lecture, ainsi qu'un modèle temporel des divers processus impliqués dans cette activité. Nous en donnons ici une version simplifiée.

Lorsqu'un mot est perçu par l'oeil, la majorité des chercheurs estiment que les récepteurs de la fovea transmettent un signal aux régions occipitales des deux cortex (Barber et Kutas, 2007, 104). Il s'agit des mêmes aires que celles qui perçoivent les visages, des objets, etc., et où s'opère l'analyse des traits visuels de l'input, environ 100 ms après l'exposition du mot à la vue du lecteur (Salmelin et Kujala, 2006, 520). Toutefois, 50 à 70 ms plus tard (Barber et Kutas, 2007; Dehaene, 2007), l'activité bascule entièrement dans la région occipito-temporale ventrale de l'hémisphère gauche, aussi appelée « la région de la forme

visuelle des mots », et dont la fonction a été mise à jour par Dehaene *et al.* (2002) à l'aide de l'IRMf.

Cette région présente un intérêt particulier pour l'étude des processus de lecture en ce qu'elle ne s'active pas lorsqu'on *entend* un mot, mais uniquement lorsqu'on le lit (Petersen *et al.*, 1989). Elle représente probablement la seule zone du cerveau qui soit propre à la lecture. Par ailleurs, elle fait partie d'un ensemble plus vaste, chargée normalement de la reconnaissance des objets, des visages, etc. (Dehaene, 2007, 111), qui collecte les informations dans les deux hémisphères pour analyser et reconnaît l'input visuel, et dont une partie se serait recyclée pour se spécialiser dans la reconnaissance des mots [42].

Dans la région occipito-temporale ventrale s'opère une reconnaissance de l'identité abstraite des lettres (Ferrand, 2007, 48), c'est-à-dire que, quelle que soit le type, la taille ou la couleur de l'écriture utilisée pour transcrire le mot reconnu, cette région est capable de réduire ces variations en une « image » abstraite. Par contre, sa sensibilité à l'appartenance des mots au lexique de la langue est plus discutée. Ainsi, Cornelissen *et al.* (2003) ont montré que cette zone de la forme ne fait pas de différence entre des mots et des suites de lettres aléatoires, tandis que Dehaene (2007, 137) rapporte que « c'est entre 150 et 190 millisecondes après l'apparition du stimulus que la région occipito-temporale gauche révèle, pour la première fois, sa préférence pour les chaînes de caractères bien formées par rapport aux suites de consonnes ou de chiffres ». On voit que l'imagerie mentale ne produit pas encore toujours des résultats convergents.

Une fois cette reconnaissance effectuée, c'est-à-dire autour de 250 ms, la région occipito-temporale redirige l'information vers de nombreuses aires du cerveau qui vont s'occuper de l'aspect sémantique et phonologique. On retrouve parmi celles-ci les aires du langage — celle de Broca, située dans le cortex frontal inférieur, et celle de Wernicke, qui se trouve dans la partie postérieure du cortex temporal inférieur) —, mais aussi un grand nombre d'autres régions, qu'ont détaillée Marinkovic *et al.* (2003) à l'aide de la magnéto-encéphalographie et dont l'activation s'effectue entre 200 et 600 ms, avec un pic à 400 (Barber et Kutas, 2007; Ferrand, 2007). Ce qui importe est que ces régions n'ont plus rien de spécifique à la lecture : elles s'activent aussi bien lorsque nous entendons des mots que lorsque nous les lisons (Dehaene, 2007, 147).

Ces diverses découvertes confirment l'existence d'au moins trois composantes distinctes pour la lecture, comme nous l'avons présenté dans la section 1.2.2 : une zone spécifique à la perception visuelle, une autre dédiée à la reconnaissance des mots et un réseau très

42. En effet, le fait qu'une zone du cerveau soit dédiée à la reconnaissance des mots, alors que le cerveau de l'homme n'a pas pu évoluer à ce point depuis les quelques milliers d'années qu'est apparue l'écriture, pose question. Pour Dehaene (2007) ou Barber et Kutas (2007), la réponse à ce paradoxe passe par l'hypothèse du recyclage neuronal.

actif et encore mal exploré qui s'occupe des divers processus de la compréhension. Cette vision théorique conduit Dehaene (2007) à dire : « Apprendre à lire consiste donc à mettre en connexion les aires visuelles avec les aires du langage ».

Conclusion

Au terme de ce parcours, nous avons retracé l'évolution de la pensée de divers hommes qui se sont successivement penchés sur la question de la lecture en L1 ou en L2. Ainsi, après une série d'études qui, au début du siècle, se concentrent surtout sur le décodage, l'étude des processus cognitifs tombe ensuite en friche durant la période behavioriste. Ce n'est que dans les années 60 que le cognitivisme revit, donnant naissance à une série de modèles explicatifs. Ceux-ci se concentrent d'abord sur le décodage (approche descendante), avant de se reporter, en un mouvement de balancier, sur la compréhension globale (approche descendante). S'ouvre alors un vif débat entre les partisans de chacune de ces deux architectures. Un débat que viendront pacifier les modèles interactifs en proposant une théorie qui combine ces deux approches. Par ailleurs, ils se font plus explicites, posant les fondements des modèles de la génération suivante, qui sont implémentables, c'est-à-dire, suffisamment explicites pour qu'une machine puisse les utiliser.

Ces modèles récents, qu'ils soient connexionnistes ou à deux routes, se caractérisent par un même état d'esprit : on limite désormais les particularismes de la lecture à la reconnaissance des mots. Cette restriction s'explique par la conviction de plus en plus répandue selon laquelle les réseaux cognitifs qui permettent de comprendre le langage sont les mêmes pour l'écrit que pour l'oral. Toutefois, il nous semble également que ce type de formalisme éprouverait des difficultés à passer au niveau de la phrase, étant donné qu'il spécifie actuellement l'ensemble de la connaissance disponible (mots, phonèmes,...), mais que cette tâche apparaît peu réalisable au niveau des phrases d'une langue. Par ailleurs, ces deux types de modèles se trouvent actuellement en compétition afin de déterminer lequel d'entre eux correspond le mieux aux structures cérébrales effectives de l'homme. Ils ont fait appel, pour trancher cette question, à un ensemble de techniques d'imagerie mentale, qui révèlent bien des choses sur les processus cognitifs de la lecture. Bien que la question ne soit pas encore complètement tranchée, le modèle à deux routes apparaît comme le meilleur candidat en termes de proximité structurelle avec les circuits cérébraux humain.

Cette longue section, qui peut paraître assez éloignée de la problématique principale de cette thèse, nous a toutefois éclairés sur la manière dont la lecture a été considérée. Rapportant cela à la lisibilité, on peut avancer l'idée que les critiques faites à la lisibilité dans les années 80 dont nous allons parler dans le prochain chapitre doivent être rapportées au cadre théorique particulier à cette époque. Aujourd'hui, un certain recul a été pris par rapport à

cette vision et l'importance des indices textuels, de la cohérence et d'autres facteurs de haut niveau a été relativisée en faveur de processus de plus bas niveau, fortement ancrés dans une dimension lexicale. Par conséquent, il nous semble que ce parcours historique autorise à penser, plus que jamais, que les formules de lisibilité et leur approche principalement syntaxico-lexicale de la difficulté des textes restent pertinentes, et ce, particulièrement dans le cadre des langues étrangères.

Dans la suite de ce chapitre, nous refermons la description des processus cognitifs de la lecture pour envisager un aspect plus directement lié à notre problématique : la place accordée au texte, littéraire ou non, au sein des différentes méthodologies de FLE. Celle-ci a évidemment été fortement marquée par les discours théoriques sur la lecture que nous venons de décrire, ce qui explique sa place au sein de ce chapitre.

1.3 La place de la lecture et du texte écrit dans les différentes approches didactiques en FLE

Aujourd'hui, le texte écrit a retrouvé une place centrale dans l'enseignement des langues étrangères. Retrouvé puisque, s'il a longtemps occupé une place de choix au sein de la première approche de l'enseignement des langues étrangères, qu'on désigne sous le terme de « traditionnelle », il a ensuite été délaissé au profit de la langue orale. Cependant, au terme de ce purgatoire qui s'achève avec la publication du CECR [43] (Conseil de l'Europe, 2001), la lecture est revenue en force, enrichie d'apports issus de diverses disciplines, au premier rang desquels figurent la psychologie cognitive et l'analyse du discours.

1.3.1 La méthodologie traditionnelle (MT)

La méthodologie [44] traditionnelle, issue de la méthodologie d'enseignement du latin et du grec, a été adaptée aux langues vivantes étrangères (LVE) à la Renaissance. Abandonnée peu à peu au début du XX^{e} en France, elle a été employée jusqu'à une époque récente dans certains pays tels que la Russie ou le Japon. Elle vise à former des lecteurs capables d'appréhender les grandes œuvres de la littérature classique en langue étrangère. Dès lors, le texte écrit et littéraire sert de support à l'apprentissage de la grammaire et du lexique, au

43. Dans la suite de notre propos, nous éviterons de surcharger le texte en désignant le Cadre européen sous le sigle CECR.

44. Dans cette thèse, nous respectons la distinction faite par Puren entre une *méthode* : « un ensemble de procédés et de techniques de classe visant à susciter chez l'élève un comportement ou une activité déterminés » (1988, 16) et une *méthodologie* : « un ensemble cohérent de procédés, techniques et méthodes qui s'est révélé capable, sur une certaine période historique et chez des concepteurs différents, de générer des cours relativement originaux par rapport aux cours antérieurs (...) » (1988, 17). De plus, le terme *approche* est ici considéré comme équivalent de *méthodologie*.

travers de l'activité phare : la traduction. C'est d'ailleurs pour cette raison qu'on dénomme également cette approche « grammaire-traduction ».

Cette prédominance du texte littéraire s'accompagne des deux postulats suivants : d'une part, la progression doit être « orientée fondamentalement sur la progression des textes à lire et à traduire » (Vlad, 2006, 26). D'autre part, la langue enseignée est celle, littéraire, que contiennent ces mêmes textes. La MT considère qu'il existe « une langue « normée » et de qualité, celle utilisée par les auteurs littéraires qui devait être préférée à la langue orale et imitée par les apprenants. » (Seara, 2002).

Toutefois, cette manière d'aborder le texte conduit à négliger la pédagogie de la lecture, comme le dénonce Lehmann (1994, 85) :

> *On fait toujours du texte écrit la source ou la ressource de toute leçon, mais c'est pour mieux l'étouffer, en visant beaucoup plus l'accumulation de connaissances sur la langue que le développement de la capacité de lecture.*

Il faut préciser, à la décharge de la MT que les avancées psychologiques sur le processus de lecture n'en sont encore qu'à leurs balbutiements en cette fin du XIX^{e} siècle. C'est ainsi que la MT a pu considérer la lecture comme « une traduction mentale instantanée de la langue étrangère à la langue maternelle » (Puren, 2008, 149).

1.3.2 La méthodologie directe (MD)

La prédominance de l'écrit a été peu à peu remise en question. Dès le XIX^{e} siècle, Félix Gouin défend, à travers son « approche naturelle », l'importance d'enseigner aussi bien l'oral que l'écrit, et précise que l'oral doit toujours précéder l'écrit dans l'apprentissage. Avec l'apparition de la méthodologie directe au début du XX^{e}siècle [45] et sous la pression de nouveaux besoins engendrés par l'accroissement des échanges économiques, politiques, culturels et touristiques, les langues ne sont plus conçues uniquement comme des vecteurs de la culture, mais aussi comme des outils de communication. L'enseignement d'une langue étrangère doit viser à une maîtrise pratique de la langue, ce qui conduit à une dévalorisation de l'écrit, du moins aux niveaux débutants.

En effet, la MD combine trois méthodes : directe, orale et active, dont la seconde recommande l'utilisation exclusive de la langue orale, sans avoir recours à l'intermédiaire de l'écrit. Il s'agit d'enseigner à livres fermés : les textes — fabriqués, adaptés ou authentiques — sont lus par le professeur et commentés oralement ou sont destinés à la récitation,

45. Puren (1988, 94) fait remonter l'adoption de la MD en France à l'Arrêté du 31 mai 1902 qui impose brutalement cette méthodologie pour l'enseignement des langues étrangères.

laquelle permet de s'assurer de la bonne prononciation des nouveaux mots. On ne peut donc guère parler de lecture à ce stade.

Toutefois, dans les niveaux supérieurs, la lecture reprend ses droits : d'une simple lecture-reprise, telle qu'elle est pratiquée au niveau débutant, elle devient lecture expliquée, où le professeur éclaire le texte de son savoir, et tend vers la lecture cursive, c'est-à-dire une autonomisation de l'élève. L'écrit est réintroduit en classe et est même considéré comme le moyen optimal d'intégrer les différentes activités d'apprentissage : l'enseignement de la phonétique, du lexique, de la compréhension orale / écrite, de la production orale / écrite, ainsi que la découverte de la littérature et de la civilisation (Puren, 1988, 171). En outre, il ne s'agit pas systématiquement de textes littéraires : l'usage de divers types de textes — journaux, revues, cartes géographiques, etc. — est envisagé.

Ce manque de continuité entre les niveaux débutants et les niveaux supérieurs peut s'expliquer par le fait que « c'était la première fois dans l'histoire de l'enseignement des LVE que se posait réellement le problème de l'articulation entre l'enseignement de la langue et l'enseignement de la civilisation étrangères » (Puren, 1988, 180). Confronté à cette problématique, la MD cherche à évoluer, en étendant ses principes aux niveaux supérieurs, tout en réintroduisant un objectif culturel à côté de l'objectif pratique. Cependant, dans le concret, on assiste plutôt au retour des textes littéraires dans les niveaux inférieurs, au détriment d'autres types de textes. Le point de vue humaniste à la base des MT remporte une victoire sur le point de vue utilitaire.

1.3.3 La méthodologie active (MA)

Cette évolution a été interprétée diversement : soit comme un retour inavoué à la méthodologie traditionnnelle (MT), soit comme une sorte de synthèse entre la MT et la méthodologie directe (MD). C'est d'ailleurs ainsi que l'on considère la méthodologie qui a suivi la MD, la méthodologie active (MA), laquelle constitue une forme de théorisation des évolutions connues par la MD et un assouplissement des méthodes directes et orales. Elle conduisit au développement de l'explication de textes à la française que résume Puren (2008, 144) :

> *Il s'agit, par un parler sur le texte qui assure l'entraînement langagier sur les formes linguistiques qu'il introduit, d'extraire du texte de nouvelles connaissances culturelles et de mobiliser des connaissances culturelles antérieures ou fournies en temps réel par l'enseignant, ou encore recherchées par les apprenants [...].*

Ainsi, malgré l'usage de l'oral en classe, la MA, comme la MD avant elle, fait la part belle à l'écrit, d'autant qu'elle le considère comme dépositaire des formes linguistiques à trans-

mettre. Toutefois, l'explication de textes tend à noyer le lecteur sous les connaissances nécessaires à l'interprétation, transformant la lecture en prétexte à des apports culturels et linguistiques, comme l'avait fait la MT auparavant, omettant qu'il existe plusieurs voies d'entrée dans un texte et plusieurs objectifs à la lecture.

1.3.4 La méthodologie audio-orale (MAO)

Aussi appelée structuro-behavioriste, cette méthodologie a vu le jour aux États-Unis, après 1945. Elle repose à la fois sur les théories behavioristes en psychologie et sur les travaux de linguistique structurale de Bloomfield. Des premières, elle adapte le schéma stimulus-réponse-renforcement au moyen d'exercices de drill qui assurent la rencontre répétée de structures linguistiques et leur assimilation sous la forme d'automatismes ; des seconds, elle retire la vision d'une langue susceptible de variations paradigmatiques et syntagmatiques, ce qui conduit à privilégier les exercices de substitution et de transformation.

Une seconde conséquence de ces emprunts à la linguistique structurale est la primauté de l'oral, qui est vu comme le mode authentique de la communication humaine : l'écrit n'en est qu'une transcription. On assiste dès lors à un renversement (Puren, 1988, 238) :

> *Il y a donc inversion du rapport écrit / oral dans la stratégie didactique : alors que dans la MA [et la MD] c'est le texte écrit qui sert aussitôt d'appui aux productions orales des élèves, dans la méthodologie audiovisuelle [et audio-orale], c'est à partir de leur compétence orale précédemment acquise que s'organisent leurs activités de compréhension et d'expression écrites [...].*

Pour Vigner (1979, 6), l'une des raisons pour laquelle l'oral est tellement favorisé par la linguistique de l'époque tient à l'incapacité de cette dernière à prendre en considération des unités de dimension supérieure à la phrase, qui sont particulièrement utiles pour la description du texte, et donc de l'écrit.

En conséquence de ces bouleversements théoriques, lorsqu'un texte est abordé, il s'agit surtout d'en pratiquer une lecture orale, au cours de laquelle le professeur aura soin de corriger la prononciation avant de préparer l'apprenant à accéder au sens au moyen de questions de compréhension. Cornaire (1999, 5) estime toutefois que « la lecture n'était rien de plus qu'un autre type d'exercice systématique (drill), et une façon plus ou moins avouée de renforcer l'oral, sans volonté de préparer l'étudiant à "lire le sens" du message » [46]. Grabe (1991, 376) s'accorde également à dire que la lecture en langue étrangère était alors considérée comme

46. Notons que cette prédominance de l'oral ne va pas de pair avec une vision communicative de la langue, comme le souligne Girard (1995, 6) : les méthodes audio-orales ou audio-linguales (...) comportaient surtout des batteries d'exercices d'entraînement systématique au détriment du développement de la communication dans des situations de la vie courante, (...) « ».

un support à la prononciation ou un moyen d'observation de la grammaire et du vocabulaire. Le problème de la méthodologie audio-orale (MAO) tient à l'acceptation d'un postulat non fondé — sur lequel la nous reviendrons par la suite — et dénoncé par Lehmann (1994, 84) : « tout se passe comme si un certain degré de maîtrise de la communication orale garantissait, dans la foulée, une maîtrise comparable de la compréhension écrite ».

1.3.5 La méthodologie audio-visuelle (MAV)

Dominante en France dans les années 60 et au début des années 70, la méthodologie audio-visuelle (MAV) repose sur une théorie linguistique conçue par Petar Guberina, la méthode structuro-globale audio-visuelle (SGAV) (cf. Guberina, 1984), qui a été adaptée en une méthodologie pour le FLE en collaboration avec Paul Rivenc. Le structuro-globalisme continue à limiter l'analyse linguistique au niveau de la phrase et reconduit l'idée de la primauté de la langue parlée, éléments tout deux déjà présents dans la MAO. Il s'agit, comme son nom l'indique, d'un structuralisme qui se veut plus englobant que ceux qui l'ont précédé puisqu'il s'agit d'analyser l'ensemble des facteurs de la situation de communication, à savoir « la situation (réelle ou dans la pensée), la signification intellectuelle et affective, tous les moyens sonores, les moyens lexicologiques et non lexicologiques, l'état psychologique des intervenants et leur co-action réciproque, leur perception et leur production satisfaisantes de la parole » (Guberina, 1984, 96).

Au niveau des méthodes utilisées dans l'enseignement, la MAV hérite de la MAO l'importance de la langue orale et de la MD le recours privilégié à la langue enseignée, à cette différence près que le cours ne repose plus sur ce que Seara (2002) appelle les « pitreries » du professeur, mais se base sur l'association de documents sonores (dialogues oraux) déjà présents dans la MAO avec un support visuel. Ce support peut consister en des diapositives, des films fixes, des dessins ou des figurines en papier floqué pour tableau de feutre, etc. et vise deux objectifs : traduire en images un contenu sémantique et expliciter les aspects contextuels (gestes, attitudes, environnement, etc.) de la situation d'énonciation des documents sonores. En effet, s'inspirant de la théorie de la Gestalt, Guberina estime que l'apprentissage d'une langue étrangère s'opère au travers des différents sens, en particulier de l'oreille et de la vue, et vise à la compréhension du sens global de la structure.

Ces différents postulats, ainsi que la subordination de l'écrit à l'oral continuent de réduire la lecture à un support pour la pratique de l'oral, comme le décrit Cornaire (1999, 6) :

> *L'entraînement à la lecture se fait en enseignant le rythme, l'intonation ; en apprenant à l'étudiant à ménager des pauses, à faire des liaisons, à placer l'accent tonique, etc. L'imitation juste de l'intonation et du rythme prime sur*

> *la compréhension du message, ce qui revient à dire que l'écrit est sacrifié à la langue parlée et plus particulièrement à la prononciation qui demeure l'élément essentiel de l'enseignement d'une langue étrangère.*

Dabène (1981, 216) ajoute : « aucune attention particulière n'est accordée aux stratégies de compréhension du sens, ni à la spécificité des situations de lecture en langue étrangère ». De plus, la question du seuil (Alderson, 1984) n'étant pas encore à l'ordre du jour, la MAV n'envisage pas la possibilité de transférer des compétences de lecture de la L1 vers la L2. Sans doute encore trop dépendante des études sur les processus de lecture en L1, elle aborde donc la question de la lecture en L2 à la manière des premiers apprentissages en L1.

1.3.6 L'approche communicative

L'approche communicative, qui se développe à partir des années 70, constitue une réaction envers les méthodologies audio-orale et audio-visuelle. En effet, ces dernières, à l'instar de la MD, peinent à organiser les contenus des niveaux intermédiaires et avancés selon la même logique que le niveau débutant (Allen et Widdowson, 1974, 2). De plus, diverses avancées en linguistique de l'énonciation, en analyse du discours ou en pragmatique ont mis l'accent sur diverses faiblesses de la MAV : oubli des fonctions émotive, phatique et conative dans les dialogues proposés aux étudiants ; passage à une unité supérieure à la phrase sans explicitation des procédés discursifs à mettre en œuvre ou encore omission des fonctions communicatives associées aux structures linguistiques (notion d'actes de parole...). Enfin, l'enseignement des LVE doit à ce moment répondre aux besoins nouveaux d'un public multiforme : migrants, scientifiques, étudiants étrangers, etc. qui désirent utiliser la langue en fonction d'objectifs spécifiques.

Ces diverses raisons conduisent l'enseignement des LVE à opérer un virage théorique à deux niveaux. D'une part, l'objectif principal n'est plus la maîtrise en soi du code ou de la structure linguistique, mais bien de la compétence communicative, une notion proposée par Hymes (1991). Celle-ci désigne « la connaissance, le plus souvent implicite, des règles psychologiques, sociales et culturelles qui régissent les échanges langagiers dans les cadres sociaux d'une communauté » (Besse, 1980, 42), ce qui requiert l'alliance des connaissances linguistiques et extra-linguistiques. D'autre part, sous l'influence cognitiviste, l'enseignement s'est centré autour de l'apprenant. Dès lors, la progression des contenus est définie en fonction des besoins langagiers des apprenants, lesquels varient selon les contextes. C'est ce qu'on nomme « l'hypothèse sociologique » (Lehmann, 1994, 56).

Dans ce contexte théorique, le texte écrit retrouve une place. D'une part, au sein de la vision communicative de la langue, il est considéré « comme une modalité autonome du langage

qui permet d'assurer la communication entre des interlocuteurs » (Cornaire, 1985, 11) et mérite donc, tout autant que l'oral, d'être un sujet d'enseignement. D'autre part, suite aux demandes de spécialistes étrangers désireux d'approfondir l'aspect écrit de la langue, il est apparu que, pour beaucoup d'apprenants, la source majeure de contacts avec la langue cible s'effectue au travers de textes écrits. Carrell (1989, 225) affirme ainsi : « in many second or foreign language learning situations, reading is a (if not the) primary source of language input for learners ».

La lecture, quant à elle, est perçue comme multiple : on parle des pratiques de lecture. Coste (1978, 27) les décrit comme « des pratiques de communication mettant en œuvre une compétence complexe aux composantes multiples » : linguistique, textuelle, référentielle, relationnelle et situationnelle. Reste cependant une difficulté majeure pour la lecture des textes littéraires mise en exergue par Puren (2008, 146). L'approche communicative cherche à favoriser les échanges entre élèves de sorte que ceux-ci soient par la suite capables de communiquer en société. Or, le texte littéraire se prête mal à cette vision, car il se définit, en général, difficilement en rapport à une série de situations de communication, la lecture constituant en soi une communication différée et personnelle avec l'auteur du texte.

1.3.7 La perspective actionnelle

L'approche communicative a cherché à supprimer la « distinction entre l'agir d'apprentissage (ou scolaire) et l'agir d'usage (ou social), son activité de référence, la simulation, consistant pour l'élève à parler en tant qu'apprenant comme s'il était un usager, à y réaliser en classe des actes de parole comme s'il agissait ainsi en société » (Puren, 2008, 148). Cette optique rend difficile la pratique de certaines activités de type « scolaire », dont la lecture, souvent difficile à insérer au sein d'une situation de communication authentique. Par conséquent est apparue une nouvelle orientation qui légitimise les activités scolaires et que le CECR désigne sous le terme de « perspective actionnelle ». Elle y est décrite en ces termes (Conseil de l'Europe, 2001, 15) :

> *La perspective (...) actionnelle (...) considère avant tout l'usager et l'apprenant d'une langue comme des acteurs sociaux ayant à accomplir des tâches (qui ne sont pas seulement langagières) dans des circonstances et un environnement donnés, à l'intérieur d'un domaine d'action particulier.*

Dès lors, il est possible de réintroduire, à côté de la lecture incluse dans le cadre d'une action sociale, une lecture vue comme une tâche scolaire au sein d'un environnement particulier, la classe. Le professeur veillera à expliciter davantage les compétences en jeu pour lire tel ou tel type de texte, ainsi qu'à orienter la lecture au sein d'un projet, qu'il soit simplement la

découverte du texte en-soi ou une tâche plus globale qui nécessite des informations contenues dans les documents analysés, ainsi que le recours à divers types de lecture. Le CECR n'affirme-t-il pas qu'il « y a « tâche » dans la mesure où l'action est le fait d'un (ou de plusieurs) sujet(s) qui y mobilise(nt) stratégiquement les compétences dont il(s) dispose(nt) en vue de parvenir à un résultat déterminé » (Conseil de l'Europe, 2001, 15) ?

1.3.8 Conclusion

La fortune du texte écrit et de sa lecture a fluctué selon les méthodologies successives. Sous la MT et la MA et, en partie, la MD, le texte écrit a d'abord constitué le support le mieux à même d'assurer la cohérence de l'ensemble de la leçon grâce aux formes lexicales, syntaxiques ou culturelles qu'il contient. Ensuite, négligé au profit de l'oral dans la MAO et la MAV, il a été réduit à un vecteur pour l'entraînement de la langue orale et de la prononciation.

Bien que la lecture de textes ait été remise au goût du jour sous l'approche communicative, est réapparue une tendance, dénoncée par Besse (1980, 88), à le percevoir avant tout comme un support destiné à présenter les diverses matières à enseigner (prononciation, grammaire, culture, etc.). Toutefois, la nouvelle perspective actionelle défendue dans le CECR a rendu aux textes écrits une place centrale dans l'enseignement des LVE : ils « ne sont plus véhicules d'autres apprentissages, mais porteurs d'actions langagières diverses » (Vlad, 2006, 30). Dès lors, en se basant sur les apports de la grammaire des textes, de l'analyse du discours ou des percées en psychologie cognitive, a pu être développée une pédagogie riche où la lecture est enfin envisagée comme une finalité en soi réquérant l'enseignement à la fois de compétences linguistiques et culturelles, mais aussi de stratégies dépendant toutes du type de texte et du contexte social dans lequel la lecture s'opère.

Toutefois, au vu de la dépendance de la didactique des LVE envers la linguistique appliquée et la psychologie cognitive, il y a fort à penser que la vision actuelle de la lecture au sein des méthodologies des LVE est encore appelée à changer en fonction des découvertes qui y seront accomplies. Ainsi, Puren (2008) imagine par exemple une analyse du texte où le choix du texte, de l'objet et des outils de l'analyse peut devenir l'enjeu d'un mini-projet.

1.4 Conclusion

Dans ce premier chapitre, nous avons présenté deux questions annexes au sujet principal de notre thèse, à savoir les études sur les processus de lecture en L1 et en L2 et la place qu'a occupée la lecture dans les pratiques pédagogiques en FLE. L'objectif de ce double état de l'art dépasse la simple culture générale et présente bel et bien des connexions essentielles

avec le reste de ce travail.

Tout d'abord, cette synthèse a servi à poser une hypothèse importante qui servira à diriger notre recherche de nouvelles variables pour évaluer la difficulté des textes en L2. Celle-ci postule que les caractéristiques des textes qui ont une influence inhibitrice sur l'une ou l'autre composante des processus de lecture représentent des variables potentiellement intéressantes pour une formule de lisibilité. Par ailleurs, soulignons que ces variables ne se situent pas uniquement au niveau des aspects de la compréhension. Nous avons en effet montré, sur la base d'un ensemble de données empiriques, que les étapes de la perception visuelle et de la reconnaissance des mots avaient un impact important sur la compréhension. Il est donc justifié de prendre en compte des caractéristiques de surface telles que la fréquence des mots, leur longueur, etc.

De plus, comme les études sur la lecture présentées dans ce chapitre ont eu une influence non négligeable sur les orientations des recherches en lisibilité, le fait d'en avoir présenté une description critique permet de relativiser certaines positions adoptées par les chercheurs d'une époque donnée. En particulier, ainsi que le développera le chapitre 2, la lisibilité a subi dans les années 80 une série de vives critiques provenant essentiellement de chercheurs travaillant sur les aspects textuels et de cohérence/cohésion, aspects qu'elle ne parvenait précisément pas à prendre en compte. Or, les psychologues cognitifs de la lecture actuels considèrent que, même si ces aspects ne sont pas à négliger, leur importance a été surestimée sous l'influence des théories de Goodman (1967) et Smith (1971). Dès lors, il nous semble que les travaux actuels, qui mettent davantage l'accent sur les processus de base telle que la reconnaissance des mots, défendent l'optique selon laquelle la lisibilité reste toujours pertinente aujourd'hui.

Enfin, le détour historique sur les différentes méthodes d'enseignement du FLE et la place que la lecture a occupée au sein de chacune d'elles a montré que cette activité a aujourd'hui retrouvé une place primordiale dans les pratiques. Le texte est dès lors reconnu comme une modalité autonome du langage recouvrant des fonctions de communication et il doit être enseigné en tenant compte de cet aspect communicationnel, voire être abordé dans le cadre de tâches plus globales. Cette position récente comporte un corolaire directement relié à la lisibilité. En effet, pour mettre en valeur l'aspect communicationnel des textes, il est justifié de recourir à des textes « authentiques », qui présentent les avantages d'être inscrits dans un contexte de communication réel, d'être constitués de données linguistiques représentatives de l'usage (Cornaire, 1999, 69) et de ne pas être limités à des textes littéraires, s'ouvrant à la « paralittérature », aux articles de journaux, aux prospectus, etc. Pour Besse et Galisson (1980, 82), ce qui compte n'est d'ailleurs pas tant l'authenticité du document – par opposition à un texte construit – mais le fait que le contenu extra-linguistique du texte reste à

la portée des apprenants, sans quoi le professeur redevient un « grand prêtre » de la lecture officiel, reléguant les apprenants à un rôle d'assistés.

Toutefois, malgré ses avantages, Cornaire (1999, 70) estime que les documents authentiques tendent à présenter un grand nombre de difficultés sur un espace réduit, au risque de décourager l'apprenant. La tentation est alors grande de revenir à des textes simplifiés, plus conformes au niveau des apprenants. Une autre solution pourrait être de trouver des textes authentiques d'une difficulté adaptée. Cependant, Antoniadis et Ponton (2004, 2) expliquent :

> *Avant l'exploitation de ressources textuelles, l'enseignant procède à leur recherche. Cette étape de recherche d'information, si elle fait l'objet de nombreuses publications dans d'autres domaines, n'est pratiquement jamais traitée dans le contexte de didactique des langues et, par voie de conséquence, aucun outil adapté n'existe. Les enseignants sont donc le plus souvent contraints d'utiliser les classiques moteurs de recherche du Web proposant des critères de sélection peu en rapport avec leurs préoccupations pédagogiques à l'exception du thème et du type de source (article journalistique, document officiel, oeuvre littéraire…).*

On imagine dès lors l'intérêt d'un outil rapide permettant de vérifier la difficulté de ces documents authentiques, aussi bien pour déterminer la nécessité de retravailler le texte que pour justement s'éviter cette étape pénible en obtenant directement un texte du niveau de complexité adéquat.

C'est pourquoi nous avons cherché, dans cette thèse, à fournir l'instrument requis pour la conception de ce genre de moteur de recherche : une formule capable de prédire la difficulté de textes. Dans les chapitres suivants, nous présenterons les formules qui existent déjà pour l'anglais et le français, avant de décrire les différents choix méthodologiques préalables à la conception d'une nouvelle formule pour le FLE, nous concentrant en particulier sur les apports du TAL.

CHAPITRE 2

UNE APPROCHE PARTICULIÈRE DE LA LECTURE : LA LISIBILITÉ

> *Quelle nécessité y a-t-il d'expliquer ce qu'on entend par le mot homme ? Ne sait-on pas assez quelle est la chose qu'on veut désigner par ce terme ? Et quel avantage pensait nous procurer Platon en disant que c'était un animal à deux jambes sans plumes ? Comme si l'idée que j'en ai naturellement, et que je ne puis exprimer, n'était pas plus nette et plus sûre que celle qu'il me donne par son explication inutile et même ridicule ; puisqu'un homme ne perd pas l'humanité en perdant les deux jambes, et qu'un chapon ne l'acquiert pas en perdant ses plumes. (Pascal, « De l'esprit géométrique »)*

2.1 Définition

2.1.1 La lisibilité

Comme le dit Pascal, certains concepts s'appréhendent mieux qu'ils ne se définissent. Peut-être en est il ainsi de la lisibilité, dont la problématique nous est plus familière qu'on ne le penserait au premier abord. Quel lecteur n'a pas un jour repoussé un livre qu'il trouvait trop ardu à son goût ? Quelle mère attentionnée ne s'est pas penchée sur un livre en se demandant si son enfant serait capable de le lire ?

Cependant, puisque les sciences modernes aiment à circonscrire leurs objets d'études, nous allons nous atteler à définir ce que l'on entend derrière le terme de « lisibilité ». Il ne s'agit toutefois pas d'une tâche aisée. De manière commune, la lisibilité désigne le degré de facilité

avec laquelle un texte peut être lu, mais la couverture sémantique de ce terme, pris dans son acception spécialisée, varie d'un auteur à l'autre, en fonction de son positionnement théorique, de la perception de la lecture qui domine à son époque ou de ses objectifs de recherche. Il est pourtant possible de faire émerger une définition plus précise de la lisibilité en la distinguant de ce qu'elle n'est pas.

Tout d'abord, la lisibilité est une caractéristique du texte et non pas du lecteur. Dès lors, la « lisibilité » est à différencier de la « compréhensibilité », concept qui englobe également la composante « lecteur » du processus de lecture, décrite dans le Chapitre 1 (Boyer, 1992, 8). La lisibilité constitue donc une perspective plus réduite sur les difficultés que rencontre un lecteur confronté à un document écrit, puisque seules les caractéristiques du texte sont prises en compte comme facteurs explicatifs de cette difficulté.

Cette vision, qui ressortit à une conception classique de la lisibilité, doit toutefois être nuancée aujourd'hui. En effet, les psychologues de la lecture se refusent à parler de composantes intrinsèques au texte. Un mot n'est-il pas ardu à interpréter si son lecteur n'en connaît pas le sens et plutôt facile dans le cas contraire ? Aussi, il nous semble plus juste de dire que, si la compréhensibilité envisage la difficulté d'un texte pour un individu précis, la lisibilité s'intéresse plutôt à la complexité de ce même texte pour une classe d'individus (Boyer, 1992, 8). Au sein de ce groupe (que ce soit une classe, un niveau d'éducation scolaire ou une classe d'âge), les caractéristiques de la population sont supposées homogènes et on peut dès lors considérer les caractéristiques textuelles comme porteuses d'une difficulté intrinsèque moyenne pour cette population. On peut dire que, pour cette population, tel mot est en moyenne plus complexe que tel autre terme.

Cette première restriction effectuée, il convient d'ajouter que le terme lisibilité ne porte pas sur l'ensemble des propriétés d'un texte, mais uniquement sur les aspects linguistiques et conceptuels. L'effet que la mise en page, les caractéristiques typographiques telles que la taille ou la couleur de la police, de la couleur du fond, etc. ou encore la présence de graphiques a sur le processus de lecture ressortit à un autre domaine, dénommé « lisibilité typographique » par Richaudeau (1969) [1].

Dès lors, que reste-t-il qui relève de la lisibilité ? Deux auteurs majeurs de la discipline, Dale et Chall (1949), l'ont défini en ces termes :

> *[It is] the sum total (including the interactions) of all those elements within a given piece of printed material that affect the success of a group of readers have*

1. Notons que l'anglais possède deux mots différents pour désigner ces deux réalités : « lisibilité » se traduit *readability* et « lisibilité typographique » correspond au terme *legibility*. Un auteur francophone, Timbal-Duclaux (1985, 14), s'inspirant de cette distinction, a proposé d'employer les termes « lisabilité » pour *readability* et « lisibilité » pour *legibility*, sans que cette suggestion ne rencontre beaucoup de succès.

> *with it. The success is the extent to which they understand it, read it at a optimal speed, and find it interesting.*

De cette définition émergent trois axes d'intérêt : le niveau de compréhension du texte, son attractivité et la vitesse de lecture. Bien que ces trois aspects soient en théorie inséparables (comprendre un texte aide à l'apprécier), ils ne firent pas l'objet du même engouement. Si Flesch (1948) est connu pour avoir développé un indice « d'intérêt humain » et que l'influence des caractéristiques textuelles sur la rapidité de lecture a été étudiée par quelques psychologues (Kintsch et Keenan, 1973; Kintsch *et al.*, 1975; Richaudeau, 1984), les études sur la compréhension sont largement majoritaires. Ainsi, l'auteur de la principale formule pour le français, Henry (1975), nous dit : « Par lisibilité, nous désignons le degré de difficulté éprouvé par un lecteur essayant de comprendre un texte ». De même, pour Klare (1963), un auteur anglophone majeur, la lisibilité correspond à la facilité de compréhension due au style.

On en revient donc à la notion de compréhension, qui reste vague et doit être distinguée de la « compréhensibilité » introduite précédemment, puisqu'elle n'envisage que le versant textuel du processus de compréhension. La question est donc de déterminer quels sont les aspects du texte qui déterminent le niveau de compréhension et qui sont pris en compte par la lisibilité. La réponse a évolué au cours du 20^{e} siècle : des seuls aspects lexicaux et syntaxiques, la lisibilité englobe aujourd'hui également les aspects de cohésion et de cohérence (Mesnager, 2002), mis en évidence par les études textuelles des années 70-80.

En résumé, la lisibilité constitue un domaine qui étudie comment associer des textes à une catégorie de lecteurs, en fonction des aspects lexicaux, syntaxiques, de cohérence et de cohésion présents dans ces textes. Elle constitue le pendant de l'évaluation des compétences chez le lecteur, qui consiste à situer un ensemble de lecteurs par rapport à un test de référence.

2.1.2 Les formules de lisibilité

Une fois circonscrit le domaine de la lisibilité, il convient de distinguer en son sein deux pratiques différentes qui visent à évaluer la lisibilité d'un texte donné. Il s'agit de la **mesure** de la lisibilité et de sa **prédiction** (Klare, 1974) [2].

Mesurer revient à évaluer la difficulté de textes en se basant soit sur des jugements de lecteurs experts (éducateurs, professeurs...), soit sur les résultats obtenus auprès d'un échantillon de lecteurs représentatifs de la population pour laquelle les textes doivent être étalonnés. Ces techniques ont été régulièrement utilisées dans divers domaines (*cf.* Chapitre 4 pour

2. Cette distinction est reprise par (Schulz, 1981, 45), qui distingue cependant le jugement de l'instructeur des tests de compréhension, aboutissant ainsi sur une classification ternaire.

davantage de précisions). Cependant, elles ne constituent pas toujours une solution pratique et efficace pour les professeurs ou les concepteurs de ressources pédagogiques, qui doivent souvent sélectionner des textes dans l'urgence. En effet, le test sur échantillon est coûteux en ressources temporelles et humaines, et le jugement d'experts requiert d'avoir un panel suffisant d'experts à disposition.

C'est pour cette raison que, dès le début du 20e siècle, on a cherché à **prédire** la lisibilité de textes à l'aide de formules mathématiques : les formules de lisibilité, remarquablement décrites par Klare (1974, 64) :

> *A readability formula uses counts of language variables in a piece of writing in order to provide an index of probable difficulty for readers. It is a predictive device in the sense that no actual participation by readers is needed.*

Plus précisément, une formule de lisibilité prend classiquement la forme suivante :

$$\textbf{Lisibilité} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \tag{2.1}$$

où

Lisibilité : correspond à l'estimation du niveau de lisibilité du texte à analyser ;

α **:** est une constante (terme indépendant) ;

$X_1, \cdots, X_n$ **:** représentent des facteurs linguistiques pris en compte dans le calcul, dont l'importance relative est pondérée en fonction des poids ;

β_i **:** sont les poids associés à chacune des variables prédictives (dit aussi indépendantes) ;

Remarquons que cette équation correspond très précisément à une équation de régression linéaire, un modèle statistique qui a longtemps dominé la lisibilité.

Les divers termes de cette équation définissent donc un modèle qui comprend quatre éléments essentiels :

– Une échelle de difficulté, généralement sélectionnée en fonction de considérations pratiques (ex. : niveaux scolaires, échelle de 0 à 100, etc.) ;
– Un ensemble de variables prédictives ou facteurs $X_1, \ldots, X_n$ qui représentent les propriétés du texte prises en compte par la formule. Citons comme exemples le nombre moyen de mots par phrase, le ratio de mots-outils ou le pourcentage de mots abstraits. Ces variables sont généralement en nombre restreint pour des raisons de parcimonie, ce qui explique que toutes les dimensions du texte ne soient jamais modélisées par une seule formule. Ce manque d'exhaustivité n'est toutefois pas considéré comme problématique étant donné que certaines variables sont fortement corrélées entre elles et donc relativement redon-

dantes [3] ;

– Un ensemble de poids à estimer $\alpha, \beta_1, \ldots, \beta_n$, qu'on désigne aussi sous le terme de **paramètres**. Ceux-ci correspondent à l'importance relative des différentes propriétés textuelles modélisées et reviennent à se poser des questions du type : quelle est l'importance du lexique par rapport aux autres facteurs (par exemple, la syntaxe) dans la compréhension ? Leur valeur est habituellement estimée par maximum de vraisemblance (Hastie *et al.*, 2001), à l'aide d'algorithmes automatisés assez complexes (par ex. Newton-Raphson).
– Une collection de textes, dont la difficulté est connue, à partir desquels il est possible d'estimer la valeur des paramètres. Leur difficulté est mesurée au moyen d'un **critère** (ex. : la vitesse de lecture du texte, le résultat obtenu à un test de compréhension ou à un test de closure, jugement d'experts...), conformément à l'échelle retenue. En général, ce critère est appliqué à un échantillon de lecteurs, dont le niveau de compétence est connu, et qui sont représentatifs de la population ciblée par la formule.

Ce portrait rapide de l'approche méthodologique standard en lisibilité met en évidence ce que sont les formules de lisibilité : un outil de positionnement, relatif ou absolu, d'un texte sur une échelle graduée de la difficulté et valable pour une population donnée. Il montre également ce que ne sont pas ces formules : un modèle explicatif de ce qui fait la difficulté d'un texte à la compréhension. Ehrlich et Tardieu (1986, 403) diront d'ailleurs : « la lisibilité est liée implicitement à la compréhension mais celle-ci n'est pas définie ; elle est opérationnalisée à l'aide de formules ».

De plus, cette première approche de la lisibilité reste partielle parce que simplificatrice. D'abord, il faut savoir que si la méthodologie présentée représente la majorité des études en lisibilité durant la majorité du 20e, d'autres ont peu à peu vu le jour. Ensuite, avec l'arrivée de l'informatique, les exigences qui pesaient sur les concepteurs de formules ont évolué : de formules parcimonieuses (c'est-à-dire ne comportant que 2 ou 3 facteurs) et simples à calculer pour un humain, on est passé à des équations plus complexes, mais où le comptage des facteurs s'est automatisé. Enfin, la liste des aspects textuels qui ont été testés est longue et il est légitime de se demander si certains d'entre eux sont de meilleurs prédicteurs que d'autres.

C'est pourquoi, dans la suite de ce chapitre, nous présentons en détail l'historique de la discipline, d'abord pour l'anglais et le français L1, puis l'anglais et le français L2. La tâche est cependant ardue ! Les formules de lisibilité ayant connu un certain succès dans le monde anglo-saxon, leur nombre a rapidement crû avec l'avancée du siècle précédent.

3. Citons comme exemple, le nombre moyen de lettres par mot et le pourcentage de mots du texte absents d'une liste des mots les plus fréquents dans la langue. Ces deux facteurs sont corrélés du fait que les mots plus longs tendent à être moins fréquents dans la langue.

DuBay (2004, 2) rapporte qu'au début des années 80, on comptait déjà plus de 200 formules. Dès lors, notre historique ne peut prétendre à être totalement exhaustif sous peine d'être rébarbatif [4]. Notre objectif est ici double. D'une part, il s'agit de dresser les grandes lignes de l'évolution de la discipline, tout en mettant en valeur les grands questionnements qui l'ont secoué et les principaux résultats obtenus. D'autre part, n'ayant pas connaissance d'une synthèse récente dans le domaine, nous avons estimé nécessaire d'en produire une qui tienne non seulement comptes des apports récents, mais articule aussi les études anglaises et françaises. Par ailleurs, ce portrait chronologique des formules et des méthodologies nous permettra également de positionner clairement nos travaux par rapport à la littérature.

4. Nous renvoyons le lecteur intéressé par d'autres synthèses sur le domaine à Klare (1963, 75-80), qui a établi une liste relativement exhaustive des formules publiées pour l'anglais jusqu'en 1959. Par la suite, Henry (1975, 19-32) a repris ce tableau concernant l'anglais et l'a complété jusqu'à 1977. Il l'a également enrichi par la prise en compte d'autres langues dont le français. Enfin, Klare (1974, 1984) détaille les études entre 1960 et 1984. Nous avons également publié une version synthétique de ce chapitre et surtout centrée sur la lisibilité du français (François, 2011).

2.2 Historique

2.2.1 Les origines

Comme souvent, on peut faire remonter les premiers balbutiements d'une discipline bien loin dans le passé. Ainsi, Lorge (1944b, 544) rapporte que dès le 10e siècle av. J.-C., les talmudistes comptaient les mots et les idées dans les rouleaux, afin de diviser les lectures hebdomadaires en des portions de compréhension équivalente. Néanmoins, le désir de mesurer la difficulté de matériaux à vocation pédagogique serait apparu durant la première moitié du 19e siècle, à une époque où, comme le rapporte Venezky (1984, 24), il était possible de rencontrer des mots tels que *foppery*, *niggardly* ou *abdicate* dans des listes de vocabulaire destinées à enseigner la lecture à des enfants anglophones. Il n'est donc pas étonnant que la première variable utilisée pour contrôler la difficulté des textes ait été le vocabulaire.

Les premières expériences sur d'autres variables, telles que l'influence de la longueur des phrases ou le type de phrase, datent de la fin du 19e siècle. Ainsi, Sherman, dans son ouvrage intitulé *Analytics of Literature, A Manual for the Objective Study of English Prose and Poetry* (1893, cité par DuBay, 2004, 10-11), met en évidence la réduction de la longueur moyenne des phrases dans les textes littéraires depuis la période pré-Élizabéthaine jusqu'à son époque. Il en déduit que le langage oral est plus efficace que le langage écrit, ce qui conduit ce dernier à se simplifier en fonction des caractéristiques de l'oral. Il estime aussi que l'emploi de phrases plus courtes augmente la lisibilité des textes.

Il fallut encore attendre les années 1920 pour que les premières formules soient développées aux États-Unis. Cet intérêt pour une mesure plus objective de la difficulté des textes s'explique par l'apparition de besoins convergents en ce sens, liés à diverses évolutions sociales qu'il est utile de rappeler. Tout d'abord, tandis que l'enseignement obligatoire se généralise, une série d'élèves qui auraient auparavant arrêté l'école après les primaires, continuent dans le secondaire. Essentiellement issus de populations immigrantes, ces étudiants se révèlent plus faibles que les élèves issus de milieux plus lettrés et trouvent les textes des manuels scolaires trop complexes (Chall, 1988, 3). À la même époque, les avancées techniques dans l'industrie, la réduction du temps de travail ou l'augmentation du chômage laisse de plus en plus de temps libre pour, notamment, lire (Zakaluk et Samuels, 1996). Dès lors, les bibliothécaires, soumis à une plus forte demande, en viennent à se plaindre du manque de lisibilité des matériaux destinés aux adultes. À ces phénomènes diffus, s'ajoute un événement plus ponctuel. Durant la dépression économique consécutive au krach de 1929, le gouvernement américain mène une campagne de ré-éducation et d'enseignement pour les chômeurs et est confronté à la nécessité de disposer de documents informatifs plus lisibles (Gray et Leary, 1935, 2).

Ces nouveaux besoins sociaux surgissent à une époque où la tendance promeut l'usage d'outils scientifiques pour étudier et mesurer objectivement les problèmes d'éducation (DuBay, 2007, 5). Parmi ces outils pédagogiques, l'un d'eux va se révéler particulièrement essentiel dans la naissance de ce nouveau domaine qu'est la lisibilité. Il s'agit du *Teachers' Book of Words*, une liste de 10 000 mots anglais classés par fréquence d'utilisation, qui est publiée en 1921 par Edward Thorndike. Pour établir celle-ci, ce psychologue de l'université de Columbia rassemble un corpus de 4 500 000 mots constitué de livres pour enfants, de textes plus techniques (livres de cuisine, manuels d'histoire...), d'articles de journaux et de textes de correspondance [5]. Ensuite, de 1911 à 1921, il passe dix ans à effectuer des comptages avec l'aide de ses assistants, pour aboutir à une liste dont l'importance sera capitale.

En effet, dans l'article qui accompagne son *Teachers' Book of Words*, Thorndike (1921) pose les bases de l'utilisation de la statistique lexicale en pédagogie. Il défend l'idée que plus un mot est fréquent, plus il est adapté à un lectorat jeune et débutant. Il se pose aussi la question de la validité des estimations obtenues et de la validité du classement des mots. Ce faisant, il repère que, pour un corpus différent de celui qu'il a utilisé, le rang obtenu pour les mots les plus fréquents constituerait une estimation plus robuste que ceux des mots situés au-delà de la seconde moitié de la liste, lesquels sont davantage susceptibles de se déplacer dans la hiérarchie.

Toutefois, l'avancée la plus notable du *Teachers' Book of Words* est qu'il offre un instrument de mesure objectif de la difficulté lexicale des textes, posant de ce fait les fondations des études postérieures en lisibilité. Lively et Pressey (1923, 395) reconnaissent déjà cette importance deux ans après la parution du livre, en écrivant : « La valeur fondamentale de la contribution de Thorndike est évidente : le *Word Book* a ouvert un tout nouveau champ d'investigation ».

Le *Teachers' Book of Words* reste d'ailleurs la référence jusqu'aux années 60, quand apparaissent les premières listes de fréquences obtenues par comptage informatique (Kučera et Francis, 1967), et ce malgré certaines faiblesses évidentes. Parmi celles-ci, Thorndike (1921, 369) lui-même mettait en évidence la mauvaise estimation des fréquences de ce que Gougenheim *et al.* (1964) appellent les mots *disponibles* [6], tandis que Dale et Tyler (1934, 394) critiquaient le fait que les homonymes ne soient pas distingués.

5. Pour connaître le détail des documents utilisés, consulter Thorndike (1921, 335-340).

6. Ce concept recouvre une série de noms, plutôt de type concret, qui présentent « une fréquence faible et peu stable, [mais] qui sont cependant des mots usuels et utiles » (Gougenheim *et al.*, 1964, 145). Citons comme exemples les mots « fourchette », « dent », « coude » ou « jupe ».

2.2.2 Les premières formules (1923-1938)

Suite à ces recherches ayant pour principal objet le contrôle du lexique, apparaissent une première série de formules de lisibilité, ainsi que des études sur les relations entre divers facteurs et les difficultés à la lecture. Nous avons rassemblé celles publiées entre 1923 et 1938, période durant laquelle ont non seulement été définis les principes théoriques du domaine, mais qui se caractérise aussi par une exploration de plus en plus approfondie de la problématique. Si les premiers travaux se limitent à quelques facteurs de type lexicaux, les recherches d'Ojemann (1934) et, surtout, de Gray et Leary (1935) envisagent un grand nombre d'indices de la difficulté, qui ressortissent à des aspects textuels très divers.

Lively and Pressey : la première formule

La première formule de lisibilité est attribuée par beaucoup (Chall, 1958; Henry, 1975) à Lively et Pressey (1923). Toutefois, il ne s'agit pas encore d'une formule au sens strict, mais plutôt d'une étude sur le meilleur moyen de mesurer la charge lexicale d'un texte. Ses auteurs se sont attaqués au problème de la sélection des manuels de science pour les cours du secondaire. À cette époque, les manuels utilisés étaient trop riches en termes techniques et les professeurs de sciences passaient plus de temps à expliquer le vocabulaire qu'à transmettre des faits. Être capable d'évaluer et de limiter cette charge lexicale aurait constitué un atout.

Dès lors, Lively et Pressey (1923) expérimentent trois méthodes différentes pour mesurer la charge lexicale de textes sur la base d'échantillons de 1000 mots :

1. le nombre de mots différents, qui est révélateur de la diversité lexicale ;

2. le nombre de mots absents de la liste de Thorndike ;

3. la médiane pondérée des rangs dans la liste de Thorndike. Elle est dite pondérée, car les mots absents de la liste sont comptés deux fois, modifiant ainsi la valeur de la médiane.

Afin de tester la validité de ces trois mesures comme indices de la difficulté lexicale, Lively et Pressey les ont appliquées à quinze ouvrages et un journal dont la difficulté était plus ou moins connue. Sur la base de deux échantillons systématiques par documents, ils ont conclu que la médiane pondérée constitue la mesure la plus efficace de la charge lexicale d'un texte. Leur contribution principale reste cependant le fait qu'ils aient été les premiers à démontrer l'efficacité des approches statistiques pour évaluer la lisibilité de textes (DuBay, 2007, 40).

Vogel and Washburne : genèse d'une méthodologie

Il faut attendre Vogel et Washburne (1928) pour voir apparaître la première formule au sens propre et la mise au point d'une méthodologie qui fera ensuite florès. Ce succès n'est pas acquis sans raison, car leur étude est certainement l'une des plus importantes en lisibilité. D'une part, les deux auteurs sont les premiers à formuler une équation de régression qui prenne en compte diverses caractéristiques textuelles, dont le vocabulaire et la syntaxe. D'autre part, ils ont défendu l'emploi d'un **critère** empirique pour évaluer la difficulté des textes qui constituent le corpus d'entraînement.

Pour définir ce critère, Vogel et Washburne (1928) se sont basés sur la *Winnetka Graded Book List* qu'ils ont publiée deux ans plus tôt. Il s'agit d'une liste de 700 livres classés selon leur difficulté et leur intérêt. Ce classement a été obtenu de la manière suivante. Chacun des ouvrages a été jaugé par au moins vingt-cinq enfants parmi les 37 000 qui ont participé à l'enquête. Les compétences en lecture de chaque enfant ayant été préalablement mesurées à l'aide du *Standfort Achievement Test*, le score d'un livre équivaut à la moyenne des résultats obtenus à ce test par les enfants qui disent avoir lu et apprécié le livre en question.

Parmi ces 700 ouvrages, nos deux chercheurs en ont sélectionnés 152 qui forment leur corpus expérimental. Pour chacun d'eux, ils ont paramétrisé, avec l'aide d'une vingtaine de professeurs volontaires, les aspects textuels suivants (Vogel et Washburne, 1928, 374-75) :

1. la difficulté lexicale, employant à cet effet les trois mesures définies par Lively et Pressey (1923) ;
2. la structure syntaxique, sur la base d'un échantillon de 75 phrases ;
3. la proportion des diverses catégories grammaticales, estimée sur un échantillon de 1 000 mots ;
4. la construction des paragraphes ;
5. la structure générale du livre ;
6. les caractéristiques physiques du texte. Cette dernière dimension, qui ressortit à la lisibilité typographique, a été assez rapidement délaissée au profit des autres dimensions.

Ensuite, chacune des variables représentant l'un des cinq premiers aspects a été corrélée avec le niveau de difficulté estimé des livres, ce qui a permis de classer l'ensemble de ces indices textuels en fonction de leur capacité prédictive. Une dizaine d'entre eux ayant été conservés pour la suite de l'étude [7], Vogel et Washburne (1928) ont expérimenté diverses combinaisons de ces dix éléments, obtenant un coefficient de corrélation multiple [8] maximal de $0,845$ avec les quatre prédicteurs suivants :

7. Pour le détail de ces dix indices et la valeur de leur corrélation avec le critère, consulter Vogel et Washburne (1928, 376).

8. Il s'agit d'une mesure de corrélation définie par Howell (2008, 521) comme « La corrélation entre le

X_1 **:** score à un test de lecture ;

X_2 **:** le nombre de mots différents dans un échantillon de 1 000 mots ;

X_3 **:** le nombre de prépositions dans ce même échantillon ;

X_4 **:** le nombre de mots de l'échantillon absents de la liste de Thorndike ;

X_5 **:** le nombre de propositions simples parmi un échantillon de 75 phrases.

Ces quatre variables ont été combinées selon l'équation de régression multiple suivante, où X_1 représente le critère utilisé, c'est-à-dire le niveau moyen de compétence nécessaire pour lire le livre donné :

$$X_1 = 17,43 + 0,085\,X_2 + 0,101\,X_3 + 0,604\,X_4 - 0,411\,X_5$$

On retrouve déjà, dans cette sélection, des indices de la difficulté lexicale, de la redondance lexicale et de la complexité syntaxique, trois dimensions qui vont dominer dans la majorité des études postérieures. Par ailleurs, les variables sont combinées linéairement et leur participation respective à la difficulté du texte est pondérée via des poids. Il s'agit d'une innovation sensible pour l'époque, qui fera de cette formule, appelée *la formule de Winnetka*, le prototype des formules modernes de lisibilité.

Quelques autres pionniers

La fin des années 20 voit encore apparaître quelques autres formules, dont l'importance est moindre. Dolch (1928) propose une étude supplémentaire sur les moyens de mesurer la charge lexicale, isolant plusieurs variables intéressantes, mais sans les combiner statistiquement. Son étude se démarque surtout par une remise en question de la pertinence du pourcentage de mots différents comme moyen de comparer des textes de taille différente (voir discussion au Chapitre 3), la mise en doute de l'association fréquence-difficulté et la prise en compte du contexte de lecture dans l'évaluation de la difficulté. Dolch observe ainsi qu'un texte destiné à une lecture hors contexte scolaire demande d'être d'un niveau de complexité inférieur à un matériau étudié en classe.

Lewerenz propose pour sa part une formule assez étonnante en 1929 (cité par Klare, 1963). Sur la base d'un échantillon de 1000 mots, il s'agit de calculer le pourcentage de mots commençant par les lettres « w », « h » et « b », plus simples en moyenne, car plus fréquents dans des textes aisés à lire, ainsi que les mots commençant par « i » et « e », estimés plus ardus en conséquence de leur fréquence plus faible. Une table fournie par l'auteur permet de

critère (Y) et la meilleure combinaison linéaire des prédicteurs », c'est-à-dire l'estimation $\hat{Y}$. Elle est désignée par la lettre majuscule R, pour la différencier du simple coefficient de corrélation noté r.

transformer ces pourcentages en valeurs dont la moyenne correspond à un niveau d'éducation américain. Lewerenz semble ainsi avoir été le premier concepteur de formule à utiliser directement comme critère cette échelle particulièrement pratique, puisqu'elle correspond à l'année scolaire des enfants.

Johnson (1930) est quant à lui le premier représentant d'une préoccupation qui domine la période classique : proposer une formule aussi fiable que possible, mais qui reste simple à mettre en œuvre. Dans ce but, il défend l'efficacité d'un indice aisément dénombrable, le pourcentage de mots polysyllabiques, sans proposer de fonction qui permette d'associer ce pourcentage à un degré d'éducation précis.

L'ouverture aux adultes

Jusqu'au début des années 30, les formules de lisibilité ne sont développées que pour les enfants. Or, à cette époque, comme règne la dépression consécutive au krach boursier, le gouvernement américain investit dans l'éducation des adultes (DuBay, 2007), pour les raisons que nous avons expliquées précédemment. Diverses enquêtes révèlent alors que les capacités de lecture de l'américain adulte moyen sont limitées et que le nombre des textes non fictionnels qui lui sont accessibles est insuffisant. Parmi ces enquêtes, citons *What People Want to Read About*, publiée en 1931 par Waples and Tyler. Les deux auteurs concluent que le lecteur moyen est souvent désireux d'acquérir de nouvelles connaissances, mais se voit freiné par la complexité des textes disponibles. Dans ces conditions, il s'avère capital d'être apte à évaluer la difficulté de matériaux textuels pour des adultes aux capacités de lecture limitées.

C'est dans ce contexte que sont publiées, en 1934, deux études explorant cette nouvelle voie. La première est due à Ojemann (1934). Elle se distingue par certains progrès méthodologiques, dont le plus notable est la définition d'un critère plus strict. En effet, si Vogel et Washburne (1928, 376) avaient déjà recouru à un critère pour évaluer la complexité de leur corpus de livres, leur approche restait peu rigoureuse. En effet, attribuer à un livre donné le niveau moyen des enfants ayant déclaré l'avoir lu et apprécié comporte plusieurs faiblesses méthodologiques. Quand l'enfant a-t-il lu le livre par rapport au test de lecture qu'il a passé ? L'enfant est-il sincère dans ses réponses ? Et surtout, peut-on déduire du fait qu'un livre ait été aimé le fait qu'il ait été compris ?

Ojemann (1934) réfléchit à une forme de critère plus valide. Dans un premier temps, il étudie la distribution des aptitudes à la lecture chez les adultes au moyen de brefs extraits traitant de sujets d'éducation parentale. Il en contrôle le niveau de compréhension à l'aide de questions ouvertes et conclut qu'une proportion considérable de ces parents éprouvent des difficultés

à lire ce genre de textes. Il démontre aussi que cette forme d'évaluation est un indicateur valable du niveau de compétence à la lecture chez les adultes.

Dès lors, Ojemann dispose d'un groupe de sujets dont les résultats au test précédent sont connus. Chaque membre de ce groupe de sujets doit alors répondre à six tests supplémentaires de 15 questions portant sur des passages choisis aléatoirement parmi un total de 16 extraits. Ces 16 passages, d'environ 500 mots, constituent le corpus utilisé pour l'étude des facteurs associés à la difficulté. Ils sont étalonnés comme suit. Pour chaque passage, on sélectionne les sujets ayant obtenu un résultat entre 9 et 12 sur ce texte, ce qui représente le score moyen et est supposé dénoter une compréhension suffisante. Ensuite, on calcule la moyenne des résultats obtenus par ces sujets au test de lecture global. Cette moyenne peut alors servir de critère, autrement dit, de variable dépendante.

Sur la base de ces 16 extraits étalonnés, Ojemann (1934) éprouve l'efficacité de six mesures lexicales et de huit aspects structuraux, mais il envisage aussi l'effet de facteurs qualitatifs tels que le niveau d'abstraction des relations, l'incohérence dans l'expression, etc. Notons qu'il ne combine pas les meilleurs d'entre eux dans une formule, sans doute parce qu'il estime que les facteurs qualitatifs, impossibles à quantifier, ne peuvent être négligés. Il propose à la place une approche qui ressortit aussi à la lisibilité : le **scaling**. Il s'agit, pour un expert humain, de jauger la difficulté d'un nouveau texte en le comparant à la référence que constituent les 16 passages ordonnés en fonction de leur complexité.

Ojemann se distingue donc parce qu'il est le premier à utiliser un test de compréhension comme critère [9], mais aussi par la profondeur de ses analyses. Il remarque l'importance des facteurs organisationnels bien avant le mouvement cognitiviste, lequel a eu tendance à ignorer la finesse des études des années 30, comme le rappelle Chall (1988, 11).

La même année, Dale et Tyler (1934) font paraître une seconde étude d'importance concernant la lisibilité de textes pour adultes. Les auteurs, s'inspirant de la méthodologie de Vogel et Washburne (1928), en reproduisent les trois étapes : identifier un critère valide représentant la difficulté qu'éprouvent des adultes à la lecture d'une série d'articles ; déterminer les facteurs influençant cette difficulté, et enfin, combiner les meilleurs d'entre eux en une équation de régression linéaire.

En ce qui concerne le critère, après avoir envisagé plusieurs possibilités [10], ils retiennent finalement l'emploi d'un questionnaire à choix multiple. Ce questionnaire comporte cinq

9. On peut cependant critiquer le caractère peu représentatif des sujets choisis. Ceux-ci, alors qu'ils sont supposés représenter l'ensemble des parents, ont été sélectionnés dans une seule ville de l'Iowa, sont majoritairement de sexe féminin et d'un niveau d'éducation apparemment supérieur à la moyenne de l'état (Ojemann, 1934, 13).

10. Pour obtenir davantage de détails sur les différentes options envisagées et leurs limites, se référer à la partie du chapitre 4 qui traite de l'utilisation des tests de compréhension à la lecture comme critère.

choix parmi lesquels le lecteur doit sélectionner la conclusion qui résume le mieux l'article lu, ainsi que la conclusion la moins adaptée au texte. Afin de maintenir constant le sujet abordé et l'objectif de lecture, Dale et Tyler (1934) ont sélectionné 74 articles courts traitant de questions de santé, un domaine dont l'intérêt auprès de la majorité des groupes d'adultes a été mis en évidence par Waples et Tyler (1931, cités par Dale et Tyler, 1934, 386). Le niveau de difficulté attribué à chaque article correspond à la moyenne des résultats obtenus par l'échantillon des lecteurs pour ce fragment. On retrouve bien là une approche similaire à celle d'Ojemann (1934).

Une fois estimée la difficulté de ces 74 fragments textuels, les deux auteurs ont expérimenté 25 facteurs, qui concernent tous des caractéristiques lexicales et syntaxiques, à l'exception du nombre de pronoms personnels. Ce dernier indice repose sur l'hypothèse qu'un texte riche en pronoms personnels est plus informel et donc plus simple. Parmi ces facteurs, Dale et Tyler (1934) en retiennent finalement trois afin de définir la formule suivante :

$$X_1 = 114,4 - 9,4\,X_2 - 0,4\,X_3 + 2,2\,X_4 \pm 9 \tag{2.2}$$

où les prédicteurs sont les suivants :

- X_1 **:** le pourcentage d'adultes, d'un niveau d'éducation équivalent à la fin du primaire qui, subissant un test de compréhension similaire à celui décrit ci-dessus, obtiendrait pour ce texte un résultat dénotant une compréhension suffisante ;
- X_2 **:** le nombre de mots techniques différents dans un échantillon dont la taille avoisine les 400 mots ;
- X_3 **:** le nombre de mots difficiles, mais non techniques, différents dans ce même échantillon ;
- X_4 **:** le nombre de propositions indéterminées [11] dans cet échantillon.

Celle-ci correspond à la combinaison la plus efficace des différents facteurs. Son coefficient de corrélation multiple ne s'élève qu'à 0.51, mais il faut prendre en compte, à sa décharge, le fait qu'elle repose sur un nombre supérieur de textes, ce qui tend à diminuer la valeur de R.

McClusky : la vitesse de lecture comme critère

L'étude de McClusky (1934), parue la même année, comporte quelques innovations intéressantes, malgré que la portée de ses conclusions soit grevée par la petitesse de l'échantillon

11. D'après Dale et Tyler (1934, 395), une proposition indéterminée est une proposition qui commence par *what*, *whatever*, *where*, *wherever*, *when* ou *whenever* et ne dénote pas une simple interrogation, mais implique une indétermination temporelle ou spatiale.

utilisé. Parmi celles-ci, la plus notable concerne le critère utilisé : McClusky (1934, 277) estime que la vitesse de lecture convient mieux qu'un score de compréhension, pour autant que la lecture soit associée à une tâche de compréhension subséquente.

Il évalue ainsi, à partir d'une population de 30 étudiants, le temps moyen de lecture de six extraits (de 580 à 919 mots) issus de six domaines différents (fiction, science politique, économie, sociologie, psychologie et physique). Ensuite, il observe les relations que ce critère entretient avec quatre variables :

– Le nombre d'idées par 100 mots ;
– Le nombre moyen de lettres par mots ;
– Le nombre moyen de mots par phrase ;
– Le nombre de mots techniques et de noms référant à des objets communs.

Parmi ces facteurs, deux d'entre eux présentent un certain intérêt pour leur aspect novateur et parce qu'ils relèvent de la dimension sémantique des textes. Il s'agit du nombre d'idées que charrie le texte, qui opérationnalise les réflexions d'Ojemann (1934), et du nombre de noms référant à des objets communs. McClusky (1934, 281) défend ici l'idée, poursuivie par Lewerenz (1935), Flesch (1950) ou Henry (1975), que les textes complexes comportent davantage de noms dont le sens est abstrait et général et moins de termes faisant référence à des objets concrets et familiers.

Gray and Leary : les défricheurs systématiques

Cette première période des recherches en lisibilité culmine avec l'étude de Gray et Leary (1935), que Klare (1963, 48) considère comme la plus exhaustive de l'histoire de la lisibilité jusque-là. Dans leur ouvrage, *What Makes a Book Readable*, les deux auteurs s'attachent à identifier ce qui rend un livre accessible pour des adultes aux capacités de lecture limitées. Pour ce faire, reprenant la méthodologie développée par Vogel et Washburne (1928) et affinée par Ojemann (1934), ils sélectionnent 48 passages d'environ 100 mots issus de textes de fiction, de magazines et de journaux [12]. La difficulté de chaque extrait est ensuite évaluée à l'aide de tests de compréhension et la collaboration de près de 800 lecteurs.

Parallèlement, Gray et Leary enquêtent auprès d'éditeurs, de bibliothécaires et d'autres experts afin d'établir une liste de 289 facteurs potentiels, organisés selon quatre catégories (Gray et Leary, 1935, 26-32) :

1. le contenu (thèmes abordés, nature du sujet, unité du contenu...) ;
2. le style de l'expression (lexique, syntaxe, effets stylistiques...) et la présentation (style de la présentation, attitude de l'auteur...) ;

12. Il s'agit ici d'un corpus généraliste et non plus de textes techniques, à la différence des précédentes recherches sur les adultes (cf. Dale et Tyler, 1934; Ojemann, 1934).

3. le format (taille du livre, typographie, format, présence d'illustrations...) ;
4. l'organisation (division en chapitres et en paragraphes, présence d'en-têtes, de diverses tables de références...).

Ces catégories sont ordonnées selon leur importance relative qui a été évaluée d'après l'avis de ces mêmes juges.

Ainsi, non content de proposer l'étude la plus exhaustive de l'époque quant au nombre et à la diversité des variables considérées, Gray and Leary ont mis à jour quatre dimensions textuelles d'où proviendraient les difficultés de lecture. Il s'agit d'une vision fine de la lisibilité qui prend en compte des aspects tels l'organisation des idées et leur cohérence, des éléments déjà mis en exergue par Ojemann (1934), et que les psychologues cognitifs des années 70 et 80 crurent oubliés des chercheurs en lisibilité. Cette méconnaissance de la richesse des premiers travaux en lisibilité s'explique sans doute par le fait que ces constatations restèrent à l'état d'idées et ne donnèrent pas lieu à des analyses empiriques.

En effet, Gray et Leary, tout en reconnaissant l'importance des trois autres catégories, se concentrent uniquement sur les facteurs stylistiques, parmi lesquels ils retiennent 82 variables (cf. Gray et Leary, 1935, 98-99). Parmi celles-ci, dix-huit ne sont pas paramétrables de manière fiable et vingt autres représentent des phénomènes trop rares pour autoriser une analyse statistique. Au final, il ne reste que 44 facteurs dont les niveaux de corrélation avec le critère sont étudiés. Une vingtaine de variables explicatives se révèlent significatives, parmi lesquelles Gray et Leary en combinent cinq en une équation de régression linéaire dont la corrélation multiple atteint $0,6435$. Cette équation s'exprime de la manière suivante (Gray et Leary, 1935, 138) :

$$X_1 = 3,774 - 0,01029\,X_2 + 0,009012\,X_5 - 0,02094\,X_6 - 0,03313\,X_7 - 0,01485\,X_8 \quad (2.3)$$

où les variables retenues sont :

X_1 **:** le score moyen qu'un groupe d'adultes aux capacités de lecture limitées obtiendrait à un test de compréhension portant sur ce texte ;

X_2 **:** le nombre de types [13] absents de la liste de Dale (1931) (qui en comprend 769) ;

X_5 **:** le nombre de pronoms personnels (1re, 2e et 3e personnes confondues) ;

X_6 **:** le nombre moyen de mots par phrase ;

X_7 **:** le pourcentage de mots différents ;

X_8 **:** le nombre de compléments introduits par une préposition.

13. Le terme « type » désigne la forme abstraite d'un mot qui regroupe ses différentes occurrences au sein d'un texte. Le nombre de types dans un texte correspond dès lors au nombre de mots différents.

Il apparaît donc que les deux auteurs, partis pour considérer un grand nombre d'aspects textuels, y compris cognitifs et organisationnels, se sont finalement rabattus sur un programme moins ambitieux. Leur étude conserve toutefois une grande importance pour son exhaustivité théorique. Gray and Leary ont considéré un grand nombre de facteurs potentiels, dont certains seront exploités dans les recherches postérieures.

Morriss et Halverson : les racines oubliées de la lisibilité cognitiviste

Avant de passer à la période suivante, retenons encore les travaux de Morriss et Halverson [14] qui essaient justement de prendre en compte les aspects liés au contenu. L'étude, qui aurait débuté autour de 1935 (Chall, 1958, 25) et dont les résultats sont rapportés dans un manuscript resté non publié et daté de 1938, apparaît très novatrice par les idées qu'elle avance :

> *They postulated that difficulty in understanding what is read depends on the author's choice of words and the complexity of the relationship of ideas. Words cannot be considered easy or hard in themselves, but only as they are used to convey easy or hard ideas. Ease or difficulty of words, therefore, must be based on the ways in which they are used in context.* (Chall, 1958, 25)

Le mot ne peut donc être simplement considéré au travers de sa forme graphique. Il faut prendre en compte les variations de sens selon le contexte, puisqu'il est évident que le verbe « être » est considérablement plus abordable que le nom, pour lequel il convient encore de distinguer entre l'« être » comme entité concrète qui existe et l'« être » comme notion philosophique abstraite. Par ailleurs, Morriss et Halverson insistent également sur une autre problématique essentielle en lisibilité, à savoir qu'une suite de mots simples peut encoder un ensemble complexe d'idées, notamment lorsque la charge inférentielle entre les différentes propositions d'un texte est importante.

Restait à mettre ces brillants principes théoriques en pratique, et c'est là que se situe l'essentiel du défi. Dans ce but, Morriss et Halverson définissent une méthode reposant sur les mots porteurs de contenu. Ceux-ci sont répartis en quatre classes qui respectent la chronologie naturelle d'apprentissage. Ainsi, les membres de la classe I concernent les expériences fondamentales de la vie que tout être humain connaît (ex. : père, eau, maison, blesser, manger...), tandis que la classe IV est réservée à des termes abstraits acquis seulement au terme d'une éducation avancée (ex. : platitude, concave, culminer...).

14. Notons qu'un désaccord existe autour du nom de cette dernière chercheuse, également citée sous le nom de Holversen. Leur manuscrit n'ayant pas été publié, nous n'avons pas pu avoir accès à la source afin de trancher cette question.

Toutefois, les deux auteurs n'explorent pas ces variables de manière systématique : ils ne disposent pas d'un corpus de textes dont la difficulté est connue et ne peuvent dès lors pas estimer les corrélations qu'entretiennent ces quatre facteurs avec la complexité des textes. Cette exploration systématique sera réalisée par la suite par Lorge (1939), qui obtiendra des corrélations assez élevées ($r = 0,7419$) avec son critère, mais situées dans le même ordre de grandeur que des facteurs structuraux classiques.

Conclusion

Comme nous l'avons vu, durant cette première période, les travaux en lisibilité se caractérisent par une innovation permanente touchant divers aspects de la discipline. On s'attache ainsi à définir un critère valide, à explorer un maximum de prédicteurs de la difficulté, tout en utilisant de mieux en mieux les outils statistiques. L'année 1934, qui voit apparaître les travaux d'Ojemann (1934) et de Dale et Tyler (1934) et l'emploi de tests de compréhension comme critère, marque un tournant. On peut dire que c'est à ce moment qu'est fixée la méthodologie qui fera florès par la suite. Elle consiste à rassembler un corpus étalonné à l'aide d'un critère et à calculer une équation de régression linéaire multiple sur la base d'un ensemble de prédicteurs de la difficulté.

Dans ces premières années, le domaine de la lisibilité reste méconnu. Les formules sont complexes à appliquer parce qu'elles comprennent de nombreuses variables indépendantes, dont le comptage n'est pas toujours simple. La période suivante va toutefois modifier cet état de fait.

2.2.3 La période classique (1938-1953)

Alors que les travaux de Gray and Leary et de Morriss et Halverson témoignent du désir d'envisager un nombre accru d'indices de la difficulté de nature de plus en plus variée, l'année 1938 marque le début d'une tendance qui va à rebours de ces formules détaillées. Les maîtres-mots de la période classique, qui s'étend de 1938 à 1953, sont « efficacité » et « simplicité ». Il s'agit désormais de concevoir une équation qui ne comprenne qu'un nombre réduit de variables (deux ou trois, au maximum), aisément dénombrables, tout en conservant un haut degré de fiabilité (autrement dit, un taux de corrélation multiple avec le critère qui est élevé).

Ce souci de simplicité s'explique vraisemblablement par les deux facteurs suivants. D'une part, les chercheurs se rendent compte que les variables prédictrices sont souvent intercorrélées et qu'il suffit d'en utiliser deux ou trois pour obtenir une formule fiable. D'autre part, comme ils aspirent à une diffusion plus large de leurs formules au sein de la société, il

s'avère nécessaire de les rendre plus faciles à employer. Par ailleurs, le contexte de l'époque va également jouer dans le sens d'une plus grande ouverture du domaine à la société civile. La seconde guerre mondiale rend en effet nécessaire la rédaction d'écrits compréhensibles par le plus grand nombre, afin que tous puissent participer efficacement à l'effort de guerre (Dale et Chall, 1948a, 11).

Dès lors, ce sont surtout les facteurs stylistiques de nature lexicale et syntaxique qui vont dominer dans les formules de l'époque (Klare, 1988, 16), dont les plus célèbres sont dues à Lorge (1944a), Flesch (1948), Dale et Chall (1948a) et Gunning (1952). Par ailleurs, plusieurs auteurs, dont Flesch ou Gunning, ne se contentent pas de définir des outils de mesure de la lisibilité des textes, mais disent vouloir attirer l'attention des rédacteurs sur certains facteurs, qui seraient des causes de la difficulté. Ils écrivent d'ailleurs des livres grand public où ils donnent des conseils d'écriture basés sur leurs recherches. Cet élan donnera naissance à la « writability », une discipline qui découle de la lisibilité, un peu malgré elle.

En résumé, la « writability » ou simplification de textes, consiste à réécrire un texte, dont la difficulté aura été évaluée comme trop élevée par une formule de lisibilité, en se basant sur les variables de cette même formule. En caricaturant, on peut dire que cette révision se limite à utiliser des termes plus fréquents, à augmenter le nombre de phrases simples, etc. Il s'agit là d'une extension du champ d'application des formules de lisibilité qui a fait l'objet de nombreuses critiques quant à sa pertinence (*cf.* Section 2.2.6).

Washburne and Morphett : le désir de simplicité

L'article de Washburne et Morphett (1938) est déjà représentatif de ce changement de préoccupations que nous venons de décrire. En effet, les deux auteurs, en remettant leur formule de 1928 sur le chantier, déclarent vouloir arriver à une formule plus satisfaisante et plus utile, ainsi qu'à une technique d'application plus simple (Washburne et Morphett, 1938, 355).

Concrètement, ils apportent trois modifications à leur formule précédente. D'abord, ils cherchent à étendre sa couverture en termes d'années scolaires couvertes. Or, l'équation de 1928 ne permet d'interpoler des scores de lisibilité que pour des enfants de la troisième primaire à la sixième secondaire, puisque seule cette population a servi à étalonner le corpus. L'extrapolation à des années inférieures, déjà peu fiable par nature, est d'autant plus risquée que Shuttleworth (1932, cité par Washburne et Morphett, 1938, 356) a montré que le critère utilisé, à savoir la moyenne des résultats obtenus au *Standfort Achievement Test* par les enfants qui disent avoir lu et apprécié chacun des ouvrages du corpus, est biaisé aux extrémités de l'échelle. Par conséquent, les deux chercheurs complètent leur corpus avec des textes adaptés aux enfants de première et deuxième primaire.

Ensuite, ils désirent rendre l'emploi de leur formule plus aisé au moyen d'un tableau de référence qui résumerait les connaissances nécessaires à son utilisation. Comme il n'est pas possible d'y faire figurer l'ensemble des 10 000 mots de la liste de Thorndike (1921), Washburne and Morphett envisagent de remplacer la variable « mots différents absents de la liste de Thorndike » par un pourcentage de mots peu communs. La notion de « peu commun » étant entendue comme un terme absent des x premiers termes listés par Thorndike, les auteurs ont recherché la valeur optimale pour x. Après expérimentation, ils en concluent qu'un mot ne doit pas être listé dans les 1 500 premières entrées de cette liste pour être considéré comme complexe.

Enfin, toujours dans un souci de simplification, Washburne and Morphett suppriment une variable : le dénombrement des prépositions. Ils se sont rendu compte qu'identifier l'ensemble des prépositions d'un texte est étonnamment difficile pour la majorité des gens. Malgré cette suppression, leur nouvelle formule [15], que nous détaillons ci-dessous, se révèle plus précise que la précédente, atteignant un taux de corrélation multiple de $0,86$.

$$X_1 = 1,294 + 0,00255\,X_2 + 0,0458\,X_3 - 0,0307\,X_4 \tag{2.4}$$

où les variables sont :

X_1 **:** niveau scolaire nécessaire pour lire le lire de manière satisfaisante ;

X_2 **:** le nombre de types dans un échantillon de 1 000 mots ;

X_3 **:** le nombre de mots peu communs différents (c'est-à-dire absents des 1 500 premiers mots de la liste de Thorndike) dans ce même échantillon ;

X_4 **:** le nombre de phrases simples parmi un échantillon de 75 phrases.

Notons toutefois que Washburne et Morphett (1938, 362) estiment le temps requis pour appliquer leur formule à un livre donné à quatre ou cinq heures pour un débutant. La simplification opérée reste donc très relative. Il faut attendre la formule de Lorge (1939) pour que l'entièreté du processus de conception ne soit dirigée par un impératif d'efficacité à l'usage.

Lorge : un tournant

Irving Lorge est certainement un personnage important du paysage de la lisibilité. Non content de concevoir, d'après Henry (1975, 21) : « la première formule moderne réellement employée », sans doute grâce à sa simplicité, il poursuit par ailleurs les travaux de Morriss et Halversen, en s'intéressant aux facteurs sémantiques de la difficulté. Dès 1938, il publie en collaboration avec E. L. Thorndike *A Semantic Count of English Words*, où est estimée

15. Notons qu'il existerait une première version de cette équation, due à Bergman (1936, cité par Klare, 1963, 52) où seul le terme indépendant change (il vaut $1,084$).

la fréquence des sens [16] plutôt que des formes graphiques (Rosenzweig et McNeill, 1962). En 1949, il sort une liste plus réduite : *The semantic Count of the 570 Commonest English Words*.

Par ailleurs, Lorge (1939) propose une première version de sa formule, qui est destinée aux enfants de la troisième primaire à la sixième secondaire. Comme nous l'avons précisé précédemment, la conception de cet outil est principalement dirigée par la volonté de déboucher sur une technique rapide et commode. Il recourt également à un nouveau type de critère : les 376 sélections tirées des livres II, III, IV et V des *Standard Test lessons in Reading* de McCall et Crabbs (1938, cités par Lorge, 1939). Il s'agit d'une série de manuels à l'usage des élèves de la troisième primaire à la deuxième secondaire qui renferment chacun de courts passages dont le placement a été effectué de la manière suivante :

> *Each lesson was administered to students along with the Thorndike-McCall Reading Scale (which yields grade scores). Sample sizes generally consisted of several hundred students for each lesson. To determine the grade scores for a lesson, a graph was made with a dot placed at the intersection of each student's raw score and his Thorndike-McCall grade score. A smooth curve was the drawn through the dots and a grade score assigned to each lesson raw score.* (Stevens, 1980)

Les avantages de ce critère sont multiples : le nombre important des textes (les formules précédentes se limitaient à quelques dizaines d'extraits), l'apparente rigueur de l'étalonnage [17], et surtout, des niveaux de difficulté exprimés sous la forme des niveaux scolaires américains (les *grades*), une échelle nettement plus concrète que le résultat moyen attendu à un test de compréhension. Ces divers atouts expliquent probablement l'énorme succès que ces manuels connaîtront par la suite en lisibilité.

Sur la base de ce corpus, Lorge (1939, 1944a) expérimente diverses combinaisons des facteurs suivants :

1. les cinq variables retenues par Gray and Leary (cf. Équation 2.3) ;
2. un score pondéré obtenu sur la base de la liste de Thorndike des 20 000 mots les plus courants en anglais ;
3. les quatre facteurs envisagés par Morriss et Holversen (*cf.* Section 2.2.2) ;
4. deux variables inspirées par Flesch (le nombre d'affixes et l'indice d'intérêt humain).

16. D'après Rosenzweig et McNeill (1962), la liste des sens est tiré de l'*Oxford English Dictionary*, publié en 1933.

17. Apparente, en effet, car Stevens (1980, 413), parmi d'autres, a jeté le doute sur la précision de cet étalonnage en ces termes : « *[...] they are not based on extensive testing and the grade scores they yield lack reliability and comparability.* »

Il parvient à une série de conclusions dignes d'intérêt. D'abord, il observe que le facteur prédominant en lisibilité est une mesure de la charge lexicale – quelle qu'elle soit –, ce que bien d'autres chercheurs confirmeront par la suite. Un second facteur d'importance se révèle être de nature syntaxique. Une fois ces deux aspects pris en compte, il semblerait que des variables plus sémantiques telles que l'indice d'intérêt humain ou l'âge d'acquisition du lexique n'apportent guère d'information supplémentaire au modèle statistique. Ensuite, Lorge remarque qu'il importe peu que l'on utilise un indice plutôt qu'un autre au sein d'une même catégorie, puisque ces différents facteurs sont fortement intercorrélés. Dès lors, autant privilégier les facteurs qui sont les plus simples à compter.

Lorge (1944a) propose la formule suivante, qui s'inspire des conclusions précédentes :

$$X_1 = 1,6126 + 0,07\,X_2 + 0,1301\,X_3 + 0,1073\,X_4 \tag{2.5}$$

où les variables sont :

X_1 **:** « une estimation du niveau scolaire qu'un étudiant moyen nécessaire devrait avoir pour être capable de répondre correctement et de manière complète à 75% des questions » d'un test portant sur le texte (Lorge, 1944a, 407) ;

X_2 **:** le nombre moyen de mots par phrase ;

X_3 **:** le nombre de syntagmes prépositionnels par 100 mots ;

X_4 **:** le nombre de mots difficiles différents (c'est-à-dire absents des 769 mots de la liste de Dale (1931)) par 100 mots.

Notons que ces trois variables (X_2, X_3 et X_4) seules atteignent un taux de corrélation multiple de $0,7699$ avec le critère. En considérant l'ensemble des variables listées aux points 1 et 2 ci-dessus, la corrélation ne s'élève qu'à 0.7722, ce qui semble donner raison à l'approche simplificatrice de Lorge. Notons également que Lorge (1948) propose par la suite une version corrigée de sa formule où les coefficients ont été recalculés afin d'éliminer une erreur de calcul. La version corrigée donne :

$$X_1 = 1,99 + 0,06\,X_2 + 0,10\,X_3 + 0,10\,X_4 \tag{2.6}$$

La commodité des variables utilisées et l'expression du niveau de difficulté sous la forme d'un niveau scolaire vont faire le succès de cette formule, qui sera également appliquée aux adultes – ce qui n'est pas recommandable –, et même employée dans une optique de simplification de textes.

Flesch : le succès grand public

Rudolph Flesch, un Autrichien qui émigra aux États-Unis pour fuir le nazisme, est sans doute le concepteur de la formule pour adultes la plus célèbre. Il la défend dans sa dissertation doctorale intitulée « Marks of Readable Style », publiée en 1943. Selon lui, il s'agit d'une approche radicalement neuve, qui ne se contente plus de considérer la charge lexicale, mais parvient à mesurer quelque chose d'autre (Klare, 1963, 56).

En étudiant les travaux de ses prédécesseurs, Flesch conclut que la place occupée par les indices lexicaux dans ces diverses formules est disproportionnée, du moins dans le cadre d'une formule qui vise un public adulte. Il appuie cette opinion sur les résultats de Gray et Leary (1935, 115), lesquels avaient révélé que les facteurs lexicaux sont moins corrélés avec la difficulté chez les bons lecteurs que chez les lecteurs les plus faibles. Autrement dit, si ces derniers ont effectivement tendance à buter régulièrement sur un mot inconnu ou complexe, les lecteurs les plus expérimentés n'éprouvent plus guère ce désagrément.

Confronté à ce diagnostic, Flesch décide de recourir à des indices dont le pouvoir prédictif reste constant quel que soit le niveau des lecteurs. Il estime ainsi que la longueur moyenne des phrases remplit cette condition. Par ailleurs, sous l'influence de Morriss et Halverson, il est convaincu que la difficulté d'un texte est liée à son niveau d'abstraction. Il modélise cette dimension à l'aide d'un paramètre fort simple : le nombre d'affixes, entendu comme l'ensemble des préfixes, suffixes et morphèmes flectionnels, à quelques exceptions près (cf. Flesch, 1946, 195-196). Enfin, il reprend l'idée qu'un grand nombre de références personnelles facilite la lecture. Alors que cette dimension était préalablement mesurée comme le nombre de pronoms de la 1re et 2e personnes, il la redéfinit, englobant également les noms propres, et une série des termes généralement associés à des individus, tels que « homme », « femme », « père » « mère », etc. (cf. Flesch, 1946, 196).

À partir de ces trois variables, il utilise les leçons de McCall et Crabbs comme critère et détermine la formule suivante (Flesch, 1946, 197) :

$$X_1 = -0,7502 + 0,1338\,X_2 + 0,0645\,X_3 - 0,0659\,X_4 \tag{2.7}$$

où les variables sont :

- X_1 **:** un score compris entre 1 (très facile) à 7 (très difficile) ;
- X_2 **:** le nombre moyen de mots par phrase ;
- X_3 **:** le nombre d'affixes par 100 mots ;
- X_4 **:** le nombre de références personnelles pour 100 mots.

Si l'on remplace la constante $-0,7502$ par $4,2498$, X_1 s'exprime alors en *grade* et corres-

pond au niveau scolaire moyen d'un étudiant qui répondrait correctement à 75% des items au cours d'un test de compréhension similaire à ceux proposés dans les leçons de McCall et Crabbs. Notons aussi que, comme Flesch s'est en partie basé sur les comptages erronés de Lorge pour concevoir sa formule, il convient plutôt d'utiliser la version suivante, corrigée par Lorge (1948) :

$$X_1 = 3,27 + 0,07\,X_2 + 0,07\,X_3 - 0.05\,X_4 \tag{2.8}$$

Par la suite, Flesch (1948) propose une version revue de sa première formule. Bien que l'application de cette dernière à un échantillon de 100 mots ne prenne que six minutes – on est loin des quelques heures requises pour utiliser Washburne et Morphett (1938) –, Flesch (1948, 222) veut encore aller plus loin dans la simplification. Il observe que de nombreux utilisateurs trouvent particulièrement fastidieux le comptage des affixes et hésitent parfois lors de leur repérage. De plus, il détecte des incohérences dans les prédictions de sa formule lorsqu'elle est appliquée à des textes écrits avec un style direct et vif, proche du ton de la conversation. Ce dernier aspect n'est pas suffisamment pris en compte.

Dès lors, il divise sa formule initiale en deux outils : l'un indique la « facilité de lecture » [18], tandis que l'autre offre un indice d'intérêt humain. Pour la première formule, Flesch conserve la variable « nombre moyen de mots par phrase » et remplace le décompte des affixes par le nombre moyen de syllabes par mots. Étant donné que ces deux indices sont corrélés à $0,87$, Flesch considère que dénombrer les syllabes constitue toujours une mesure du degré d'abstraction. Il obtient l'équation suivante, dont le coefficient de corrélation multiple atteint $0,7047$ (Flesch, 1948, 225) :

$$\text{Reading Ease} = 206,835 - 0,846\,wl - 1,015\,sl \tag{2.9}$$

où :

Reading Ease (RE) : un score compris entre 0 (très difficile) et 100 (équivaut à un texte au sujet duquel un enfant de quatrième primaire obtiendrait un résultat de 75% à un test de compréhension [19]) ;

wl **:** le nombre de syllabes pour 100 mots ;

sl **:** le nombre moyen de mots par phrase.

En ce qui concerne l'indice d'intérêt humain, Flesch reprend la variable « nombre de références personnelles », dont il modifie les règles de comptage. Désormais, sont pris en

18. Cette traduction de « Reading Ease » est reprise de Conquet et Richaudeau (1973, 7). de Landsheere (1982) utilise aussi les termes *note de facilité* ou *test de facilité*.

19. Cette précision révèle que cette échelle centésimale a été mise au point en multipliant les *grades* des leçons de McCall et Crabbs par dix. Dès lors, un point correspond à un dixième d'année scolaire. Toutefois, Flesch (1948, 225) nous avertit que cette correspondance ne tient que jusqu'au septième *grade* et, qu'au-delà de ce seuil, sa formule sous-estime le niveau scolaire des textes évaluées de manière croissante.

compte l'ensemble des noms comportant un genre naturel, tous les pronoms sauf le pronom neutre, ainsi que les vocables *folks* et *people*. Il ajoute une nouvelle variable : le pourcentage moyen de « phrases personnelles » [20] qui lui donne l'équation suivante (Flesch, 1948, 225) :

$$\text{Human Interest} = 3,635\,pw + 0,314\,ps \tag{2.10}$$

où :

Human Interest (HI) : un score compris entre 100 (représente là aussi à un texte dont le niveau d'intérêt humain est adapté à un enfant de quatrième primaire) et 0, qui ne s'interprète pas de la même façon que dans la formule précédente. Il marque simplement ici un passage qui ne contient ni « mots personnels », ni « phrase personnelle » ;

pw **:** le nombre de mots personnels par 100 mots ;

ps **:** le nombre de phrases personnelles par 100 mots.

L'HI, foncièrement novateur, présente toutefois le défaut de n'avoir pas été mesuré directement. En effet, le critère reste le même que pour le Reading Ease (RE), à savoir le niveau scolaire moyen des enfants ayant obtenu un résultat de 75% aux tests de compréhension portant sur les textes du corpus. Dès lors, ce que mesure réellement l'Human Interest (HI) est l'effet – au sens statistique du terme – des deux variables d'intérêt humain sur la compréhension. Notons que cet effet est faible : R ne dépasse pas $0,4306$, ce qui veut dire que la combinaison de ces deux variables n'explique que $18,54\%$ de la variance de la compréhension.

Grâce à ces deux formules, Flesch va jouir d'un large succès, qui est autant dû à ses efforts de vulgarisation qu'à la simplicité de ses outils (en cela, il est un représentant typique de ce désir de simplification qui domine la lisibilité dans sa période classique). Toutefois, parallèlement, Flesch a fait l'objet de nombreuses critiques de la part d'autres spécialistes du domaine. Il est vrai qu'il ne pèche pas par un excès de rigueur méthodologique. Ainsi, alors qu'il dit vouloir développer une formule généraliste, plutôt à destination des adultes, il se base sur les leçons de McCall et Crabbs qui ont été étalonnées pour des enfants. Flesch (1948, 222) le reconnaît lui-même et s'en explique ainsi :

> *For many obvious reasons, the grade level of children answering test questions is not the best criterion for general readability. Data about the ease and interests with which adults will read selected passages would be far better. But such data were not available at the time the first formula was developed, and they are still unavailable today.*

20. Sous ce terme vague, Flesch (1948, 223) englobe les phrases entre guillemets, les questions, les ordres, les requêtes et les autres moyens d'interpeler le lecteur, ainsi que les exclamations et les phrases grammaticalement incomplètes dont le sens doit être inféré depuis le contexte.

Par ailleurs, il n'hésite pas à étendre le champ d'application de sa formule à des populations autres que celle qu'il a modélisée, qu'il s'agisse des lecteurs ou des textes. Ainsi, il se réjouit d'avoir employé avec succès son outil dans des domaines aussi variés que les journaux, la publicité, des publications gouvernementales, des prospectus pour fermiers, du matériel éducatif pour adultes ou des livres pour enfants (Flesch, 1948, 221). On peut toutefois douter que les extraits qui composaient son corpus rassemblent une telle variété de textes, ce qui fait peser de sérieux soupçons sur la validité des prédictions réalisées dans ces contextes étendus.

Enfin, Chall (1958, 32) n'a pas tardé à mettre en doute l'aspect novateur dont se targuait Flesch. Elle soupçonne que le degré d'abstraction d'un vocable soit corrélé à sa longueur et que comme un mot plus long comprend généralement plus d'affixes, ce dernier indice ne constitue pas une meilleure mesure de l'abstraction que d'autres variables telles que le nombre de mots monosylabiques ou le nombre moyens de lettres par mots (*cf.* Chapitre 3).

On pourrait encore gloser longtemps sur les faiblesses méthodologiques de Flesch, mais la critique ne doit pas faire oublier que c'est lui qui a eu le mérite de faire sortir les études en lisibilité de leur cercle restreint et d'y intéresser de nombreux acteurs sociaux. C'est lui qui a contribué à diminuer la difficulté des histoires en première page des journaux du 16e *grade* au 11e *grade* (DuBay, 2007, 200). C'est encore lui qui proposa la première formule applicable en quelques minutes à peine, un avantage qui lui vaut d'être toujours utilisée aujourd'hui dans des logiciels comme Microsoft Word ou d'être encore enseignée dans certaines facultés de journalisme.

Dale and Chall : la rigueur classique

En réponse aux travaux de Flesch sur les adultes, Edgar Dale et l'une de ses étudiantes, Jeanne Chall, développent une formule qui restera longtemps considérée comme la plus fiable et la plus utilisée dans le milieu éducatif (Klare, 1988, 18). En effet, ces deux chercheurs ont mis au jour plusieurs défauts dans la première formule de Flesch.

Tout d'abord, ils observent que le comptage des affixes sur un même échantillon, pratiqué par plusieurs personnes et sans le recours systématique à un dictionnaire, ne conduit pas à des résultats parfaitement reproductibles. La consultation d'un dictionnaire est bien sûr possible, mais allonge considérablement le processus. Dale et Chall (1948a, 13) considèrent que le recours à un inventaire regroupant les mots « faciles » est plus précis. Il leur faut toutefois revoir la définition de « facile », car Flesch a précédemment pointé du doigt l'inaptitude de la liste de 769 mots de Dale (1931) à discriminer correctement des textes plus complexes que le huitième *grade*. La solution qu'ils envisagent alors est de recourir à une liste plus

longue, qui reprend les vocables connus des étudiants de quatrième année primaire. Celle-ci a été constituée par Dale en 1931 de la façon suivante. Sur la base d'un inventaire de 10 000 termes, Dale a testé le bagage lexical d'un échantillon d'élèves en quatrième année primaire. Il a ainsi récolté une mesure de familiarité pour chaque mot et seuls ont été retenus ceux qu'au moins 80% des élèves ont déclaré connaître, ce qui donne une liste de 3000 mots.

Revenant à la critique de Flesch, Dale et Chall (1948a, 15) estiment que le nombre de mots « personnels » n'est guère utile à la prédiction de la lisibilité des textes. D'autant que la manière dont s'effectue la paramétrisation est critiquable en raison de la relation ambiguë qu'entretiennent les noms propres avec la difficulté (sur ce sujet, *cf.* Chapitre 3). Par conséquent, ils abandonnent purement et simplement ce facteur, tout en gardant par contre la variable syntaxique désormais classique : le nombre moyen de mots par phrase.

Une fois ces deux indices sélectionnés, il reste à Dale et Chall (1948a) à définir un critère. À l'instar de Lorge et de Flesch, ils se servent des leçons de McCall et Crabbs – à défaut d'autre chose – et attribuent à chacun des 376 passages de ce manuel le niveau scolaire moyen des élèves ayant répondu correctement à 50% des questions. Corrélant ces niveaux scolaires avec leurs deux variables indépendantes, ainsi que celles utilisées par Flesch (1946) et Lorge (1944a), ils en arrivent aux mêmes conclusions que Lorge (Dale et Chall, 1948a, 17) :

> *This table corroborates Mr. Lorge's findings that a measure of vocabulary load is the most important factor in reading difficulty, and that all the measures of vocabulary are highly intercorrelated.*

L'équation de régression linéaire qu'ils obtiennent est la suivante (Dale et Chall, 1948a, 18) :

$$X_1 = 3,6365 + 0,1579\,X_2 + 0,0496\,X_3 \tag{2.11}$$

où les variables représentent :

- X_1 **:** le niveau scolaire moyen d'un élève qui répondrait correctement à 50% des questions à un test de compréhesion sur ce texte ;
- X_2 **:** le pourcentage de mots absents de la liste de 3 000 mots de Dale [21] ;
- X_3 **:** le nombre moyen de mots par phrase.

Elle obtient un coefficient de corrélation multiple de $0,70$ avec le critère. Il s'agit d'un résultat légèrement meilleur que les coefficients corrigés de Flesch (1946) et Lorge (1944a), qui atteignent tout deux $0,66$, mais équivalent à celui de la RE de Flesch (1948). En sus de

21. Cette liste, ainsi que diverses conventions de comptage, ont été publiées dans un second article (1948b) par les deux auteurs.

ces remarquables validité et fiabilité, diverses expériences de validation ont été effectuées avec succès sur d'autres textes. Celles-ci ont également permis à Dale et Chall (1948a, 18) de proposer une table de correction des résultats bruts produits par leur formule afin d'obtenir, pour les matériaux testés, une évaluation plus adéquate du niveau scolaire.

Ce dernier point constitue une nouveauté méthodologique essentielle et pourtant peu mise en évidence dans la littérature, qui revient à appliquer une fonction non linéaire sur le résultat de l'équation de régression linéaire. Dès lors, la relation entre le critère et les variables prédictrices devient elle-même non-linéaire. Cependant, Dale and Chall en restent à un niveau empirique. Le constat théorique que la relation entre la difficulté et ses indices n'est pas nécessairement linéaire ne sera explicité que presque vingt ans plus tard, par Bormuth (1966).

Largement reconnue pour ses qualités, la formule de Dale et Chall (1948a) ne sera pas pour autant exemptée de critiques. Ainsi, Benoît (1986b) désapprouve le recours à la liste des 3 000 mots de Dale qui, en 1948, date déjà de quelques années et, surtout, n'a pas été conçue pour l'enseignement secondaire, alors que la formule prétend être adéquate de la troisième année primaire à la sixième secondaire (et donc aussi pour les adultes ayant interrompu leur scolarité à l'un de ces niveaux). Notons que la question de l'ancienneté de la liste devient de plus en plus problématique avec le passage du temps (*cf.* travaux de Holmquist à la Section 2.2.4). C'est l'une des raisons qui poussent Chall et Dale (1995) à effectuer une révision de leur formule, publiée après la mort d'Edgar Dale (*cf.* Section 2.2.7).

Deux formules pour enfants

À côté de ces deux formules majeures, l'année 1948 voit aussi la publication de deux formules destinées spécifiquement à l'évaluation des matériaux pour les jeunes enfants. Elles sont respectivement dues à Dolch (1948) et à Wheeler et Wheeler (1948).

Dolch (1948), comme d'autres avant lui, a commencé par envisager diverses facettes des livres (typographie, quantité d'illustrations, niveau d'abstraction, rapport du contenu à l'expérience du lecteur, etc.) qui puissent expliquer leurs différences, avant de se rabattre uniquement sur les deux facteurs structurels classiques : la longueur des phrases et la charge lexicale. Ses travaux explorent diverses techniques de paramétrisation de ces deux aspects, sans les combiner en une formule. À la place, des standards de comparaison exprimés en niveau scolaire sont fournis pour chacune des variables.

L'étude est la première, à notre connaissance, à utiliser d'autres mesures statistiques que la moyenne ou la médiane comme indicateurs de la distribution des variables indépendantes. En effet, Dolch (1948, 246) estime, en ce qui concerne la longueur des phrases, que certains

livres ayant la même médiane ont pourtant une proportion de phrases longues bien plus élevée que d'autres. Il ne considère pas uniquement la phrase médiane (en nombre de mots), mais prend aussi en compte la phrase « longue » typique, qui correspond au 9e décile de la distribution [22]. Au niveau des facteurs lexicaux, il analyse les effets du pourcentage de mots polysyllabiques, du pourcentage de mots de plus de deux syllabes ou encore de mots absents d'une liste de 1 000 mots qu'il a constituée à partir de celle de 769 mots de Dale.

De leur côté, Wheeler et Wheeler (1948) décrivent également une série de facteurs textuels influençant l'expérience de lecture des enfants, tels que l'organisation du texte, la quantité et la qualité des illustrations, la structure syntaxique, les aspects typographiques et de format ou encore les références socio-culturelles présentes. Cependant, leur « formule » de lisibilité ne prend finalement en compte que le vocabulaire. Elle nécessite d'inventorier environ 1 000 mots issus de courts extraits choisis systématiquement toutes les 5 à 10 pages du livre à évaluer, avant d'attribuer à chacun de ces mots leur niveau scolaire, tel qu'il est défini dans le *Thorndike Wordbook of Twenty Thousand Words*. Ensuite, il ne reste plus qu'à établir le pourcentage de mots pour chaque niveau scolaire, ce qui constitue une sorte de profil lexical du texte.

Reste à interpréter ces pourcentages. Pour ce faire, les auteurs partent du postulat suivant :

> *[...] the child should know 90% of the vocabulary of the materials he uses in instructional reading, and at least 95% of the vocabulary he meets in his independent reading.*

Dès lors, le niveau d'un texte correspond à l'année scolaire pour laquelle 90% ou 95% des mots de l'échantillon sont d'un niveau égal ou inférieur. Toutefois, cette distribution des mots par niveau scolaire permet un diagnostic plus fin en fonction des objectifs d'enseignement. Wheeler et Wheeler (1948) défendent en effet une idée très importante, déjà introduite dans le chapitre 1, mais encore assez neuve à l'époque, à savoir qu'une formule de lisibilité doit être adaptable au contexte de lecture. En effet, les capacités de lecture des élèves mesurées par des tests standardisés constituent un maximum. Par conséquent, les formules conçues d'après ces tests courent le risque de sous-estimer la complexité des matériaux pour une lecture non encadrée.

22. Un décile est l'une des 9 valeurs qui divisent un ensemble de données en 10 parts égales. Autrement dit, le 9e décile correspond à l'observation qui sépare les 90 premiers pourcents de la distribution des 10% restants (Howell, 2008, 57)

Flesch et Gillie : retour sur la question de l'abstraction

Deux ans après sa célèbre formule de « facilité de lecture », Flesch se remet à l'ouvrage et décide de s'attaquer plus spécifiquement à la question de l'abstraction, en s'appuyant sur les techniques développées pour la lisibilité. Dans un article de 1950, il suggère une nouvelle mesure du niveau d'abstraction d'un texte, qui repose sur l'hypothèse que certaines catégories grammaticales sont plus fréquemment associées à un discours abstrait que d'autres. Flesch explique comment repérer et dénombrer ce qu'il appelle les « mots définis », lesquels comprennent notamment les noms communs et les noms propres possédant un genre naturel ou apportant une indication temporelle, les verbes finis, les interjections, etc. [23].

Sur la base de cette nouvelle variable et des leçons de McCall et Crabbs, Flesch développe une formule qui ne fait appel qu'au nombre de « mots définis » et exprime le degré d'abstraction du texte sous la forme d'un indicateur compris entre 0 et 100. Toutefois, l'efficacité de cet unique prédicteur est limitée ($r = -.5541$). Sans doute parce que, comme c'était le cas pour pour l'HI, l'effet mesuré par la corrélation ne représente que la perte de compréhension lorsque le texte devient plus abstrait.

Dès lors, Flesch cherche à améliorer la capacité prédictive de ce nouvel indice en le combinant avec le nombre moyen de syllabes pour 100 mots, qu'il avait déjà employé dans sa « Reading Ease ». Il parvient à l'équation suivante (Flesch, 1950, 385) :

$$R = 168,095 + 0.532\,dw - 0,811\,wl \tag{2.12}$$

où :

R **:** un score compris entre 0 (très difficile) et 100 (équivaut à un texte pour lequel un enfant de quatrième primaire obtiendrait un résultat de 75% à un test de compréhension [24]) ;

dw **:** le nombre de « mots définis » ;

wl **:** le nombre de syllabes pour 100 mots.

Bien que le coefficient de corrélation multiple de l'équation s'élève à 0.72, c'est-à-dire un peu plus haut que celui de la RE, cette nouvelle formule a très rapidement fait l'objet de vives critiques de la part de Jenkins et Jones (1951). Ceux-ci estiment que la procédure de comptage de cette nouvelle variable est trop complexe – il n'y a pas moins de seize étapes – et qu'elle requiert des connaissances grammaticales avancées. Il n'est pas donné à tout le monde de pouvoir déterminer ce qu'est un verbe fini, un pronom indéfini ou quand le verbe

23. Pour connaître le détail des normes de comptage, le lecteur intéressé se reportera à Flesch (1950, 387-388).

24. Les remarques de la note 19 sont également valables pour cette échelle, puisque ce sont les mêmes.

« être » est employé comme copule. Cette complexité est d'autant moins utile, qu'il est hautement probable que le nombre de « mots définis » soit fortement corrélé à des variables plus simples à mesurer, telles que la longueur moyenne des mots.

Flesch (1951, 69) répondra à ses détracteurs en rappelant une donnée trop souvent oubliée :

> *However, readability surely can not be defined in terms of the criterion used [...]. Rather, readability is a complex quality of written prose that relates to a variety of factors on the part of readers on different levels of age and education [...].*

Autrement dit, Flesch nous explique qu'il ne se situe plus dans la philosophie de la période classique où l'objectif est d'optimiser la valeur du R et le temps d'application en recherchant les meilleurs prédicteurs possibles. Il cherche plutôt à identifier certaines **causes** de la difficulté, ce qui présente nettement plus d'intérêt pour la « writability ».

Quelques années plus tard, Gillie (1957) revient sur cette nouvelle variable de Flesch. Dans la lignée de l'esprit classique, il cherche à en simplifier le dénombrement. Lors d'une première étape, il étudie séparement la capacité prédictive de chacune des seize règles de Flesch (1950), n'en retenant que deux. Il y ajoute une troisième : le nombre de « noms abstraits », définis comme les noms terminés par les suffixes *-ness*, *-ment*, *-ship*, *-dom*, *-nce*, *-ion* et *-y* (sauf les diminutifs). Il combine ces trois variables au sein de la formule simplifiée suivante :

$$\text{Niveau d'abstraction} = 36 + X_1 - 2X_2 \tag{2.13}$$

où :

Niveau d'abstraction : un score compris entre 0 (très abstrait) et 90 (très concret), qui correspond au résultat de la formule de Flesch (1950) ;

X_1 **:** le nombre de verbes finis + le nombre d'articles définis et de leurs noms ;

X_2 **:** le nombre de noms abstraits.

Notons que, comme l'a fait lui-même remarquer Gillie (1957, 216), la validité de cette formule ne peut dépasser celle de Flesch (1950), puisqu'elle lui sert de critère.

Gunning : l'homme d'affaires

La dernière des quatre grandes formules classiques est due à Gunning (1952). Elle s'accorde bien avec l'esprit de la période classique. Elle est simple, efficace et se veut un guide pour la rédaction. Son concepteur, Robert Gunning, était un homme d'affaires américain, ce qui explique sans doute pourquoi sa méthologie est plus lâche que celle, par exemple, de Dale et Chall. Toutefois, dans l'ensemble, Gunning respecte la procédure de développement de

l'époque dans laquelle deux facteurs structuraux sont corrélés avec le niveau scolaire des leçons de McCall et Crabbs.

Les travaux de Gunning (1952), proches de ceux de Flesch, s'en distinguent par quelques variations, dont la plus importante et la seule que nous envisageons ici concerne la mesure de la charge lexicale. L'homme d'affaires pragmatique qu'est Gunning estime que le décompte des syllabes reste pénible et, surtout, ne va pas à l'essentiel, c'est-à-dire le repérage des mots difficiles. Le nombre de mots de trois syllabes ou plus lui apparaît comme plus pertinent et plus simple à établir. Il obtient dès lors la formule simplifiée suivante (Gunning, 1952, 38) :

$$\text{Fog Index} = 0,4\,(X_1 + 100X_2) \tag{2.14}$$

dans laquelle les variables sont :

Fog Index : un score qui correspond au niveau scolaire moyen des élèves ;

X_1 **:** le nombre moyen de mots par phrase ;

X_2 **:** le ratio des noms difficiles, c'est-à-dire comportant au moins trois syllabes (à quelques exceptions près), sur le nombre total de mots de l'échantillon.

Le nom donné à la formule, l'indice Fog, vient de la mise en garde que Gunning (1952, 39) répète dans son livre :

> *If your copy tests 13 or more [at the Fog Index], you are beyond the danger line of reading difficulty. You are writing on the college level of complexity and your reader is likely to find it heavy going even though he is paying close attention.*

Autrement dit, au-delà d'un indice de 12, vous êtes dans une zone de flou, de brouillard... d'où ce nom d'indice Fog.

Notons que le nom fera florès, puisque McElroy (1953, cité par Klare, 1963, 65-66) proposera à son tour le McElroy's Fog Count. Il s'agit simplement de la somme pondérée des syllabes, qui est ensuite normalisée pour exprimer un résultat sous la forme de niveaux scolaires.

Conclusion

Avec ces deux dernières formules, s'achève la période dite classique qui, comme nous l'avons répété, se caractérise par un désir de simplicité. Lors de la période précédente, on cherchait à perfectionner la méthodologie, à explorer divers types de critères et à envisager une grande variété de facteurs prédictifs. À la période classique, l'innovation se situe essentiellement au niveau des variables prédictives : les chercheurs veulent limiter le nombre des variables indépendantes – certains iront même jusqu'à n'en utiliser qu'une seule. Par contre,

la méthodologie est très uniforme : elle repose presque toujours sur les leçons de McCall et Crabbs.

Il faut mentionner quelques limitations de ces fameuses leçons, autres que celles déjà citées. Ainsi, aussi bien Lorge (1944a), Flesch (1946, 1948) que Dale et Chall (1948a) et Gunning (1952) ont utilisé les livres II à V des leçons de McCall et Crabbs, qui ne couvrent qu'une période allant de la troisième primaire à la deuxième secondaire. Cet intervalle ne correspond pas complètement à l'échelle utilisée par ces diverses formules, laquelle s'étend de la troisième primaire à la sixième secondaire. Cela explique notamment pourquoi la qualité des prédictions se dégrade pour les niveaux supérieurs, puisqu'on se trouve en situation d'extrapolation. On peut donc douter que ce critère soit adapté aux ambitions des concepteurs des formules classiques, d'autant que d'après Stevens (1980), l'étalonnage de McCall et Crabbs est loin d'être suffisamment rigoureux pour l'usage qui en a été fait en lisibilité.

Par ailleurs, notons qu'on rencontre bien l'une ou l'autre approche dissidente au sein de ce conformisme méthodologique. Ainsi, Flesch (1950, 1951) nous semble avoir ressenti les limitations de la démarche de l'époque. Certes, les formules sont considérablement plus simples à employer, mais le nombre restreint de prédicteurs les rend moins sensibles à certains aspects textuels susceptibles d'apparaître dans des cas plus spécifiques (notamment les textes de spécialités). Il s'agit d'une faiblesse que la période suivante va chercher à pallier en développant des formules spécifiques à un type de texte ou de lecteur.

2.2.4 Perfectionnements et spécialisations (1953 à 1965)

Après que les formules de Lorge, Flesch et Dale and Chall sont parues, le nombre de travaux augmente à mesure que croît l'intérêt du public envers la lisibilité (l'armée américaine, notamment, s'intèresse de plus en plus à la question). Pourtant, l'ambition qui a marqué la période classique s'essouffle. Nombre de chercheurs abandonnent l'idée de développer une nouvelle formule qui soit universelle, et se contentent de perfectionner les formules classiques. Ainsi, une part importante des travaux de cette période sont des simples tentatives d'amélioration des quatre formules classiques les plus célèbres : Lorge, Flesch, Dale and Chall ou Gunning.

Parallèlement, apparaît une seconde tendance, mise en évidence par Klare (1963, 66-67) : la conception de formules spécialisées. On se rend de mieux en mieux compte des limites de la méthodologie statistique qui est employée, notamment au niveau de l'échantillonnage. Par conséquent, il est plus sûr, méthodologiquement parlant, de développer des formules adaptées à un sous-ensemble textuel. Il peut s'agir de textes scientifiques (Brown, 1965; Jacobson, 1965 ou Holmquist, 1968, cité par Klare, 1974), religieux (Stocker, 1971), de

textes pour les débuts des primaires (Spache, 1953; Stone, 1956), ou encore de tests à la compréhension (Forbes et Cottle, 1953). Certains chercheurs vont même se pencher sur la compréhension de l'oral (Fang, 1966; Rogers, 1962). Parmi ces formules spécialisées, certaines sont originales (Jacobson, 1965; Spache, 1953; Stone, 1956), d'autres reposent sur des formules classiques (Brown, 1965; Stocker, 1971 ou Holmquist, 1968, cité par Klare, 1974).

Enfin, on trouve quelques nouvelles formules à portée plus généraliste, où apparaissent la majorité des innovations qui vont marquer la période. La principale d'entre elles est l'automatisation progressive des formules à l'aide d'une technologie émergente : l'ordinateur. Cette tendance à l'automatisation est initiée par Smith (1961), mais il faut attendre Danielson et Bryan (1963) pour que la première formule automatique soit effectivement implémentée.

Notons par ailleurs que, de manière générale, les paradigmes théoriques restent ceux de la période classique. L'idéal conjoint de simplicité et d'efficacité est toujours de mise : les formules ne comprennent pour la plupart que deux variables.

Les révisions et simplifications

Déjà durant la période classique, certains développeurs, tels Lorge ou Flesch, ont estimé nécessaire de réviser leur formule pour en éliminer une erreur de calcul ou un défaut révélé par la pratique. Cette tendance à la révision s'accroît durant cette période de transition et se concentre autour des quatre formules classiques les plus célèbres. Elle est stimulée par les raisons suivantes :

– Tout d'abord, les leçons de McCall et Crabbs connaissent deux rééditions : en 1950 et 1961. Ce recueil de textes avait été étalonné au moyen d'une série de tests effectués sur une population d'élèves avant 1926, date de publication du manuel. Dès lors, dans les années 50-60, la langue représentée ne reflète plus le réel et la population testée a évolué. Cette réédition est l'occasion de moderniser les modèles existants et de comparer des formules entraînées sur des critères différents, mais comparables ;
– Par ailleurs, la recherche de nouvelles variables prédictives se poursuit, toujours dans une optique de simplification maximale des efforts à fournir pour appliquer la formule ;
– Enfin, comme nous l'avons déjà dit précédemment, quelques chercheurs vont adapter des formules généralistes à un type de texte particulier.

Dans la suite de cette section [25], nous envisagerons respectivement les révisions des formules de Lorge, de Flesch, de Dale et Chall et de Gunning en cherchant à rester le plus synthétique

25. Certains travaux présentés ci-après sont postérieurs à la frontière temporelle que nous avons fixée pour cette troisième période. Nous les regroupons tout de même dans cette section, pour des raisons de cohérence, mais aussi parce que leurs caractéristiques les rapprochent davantage de cette période que de la suivante.

possible.

La formule de Lorge Il y a eu peu de reprises des travaux de Lorge, sans doute parce que sa formule est la plus longue à employer du quatuor classique. La première révision est due à Barker et Stokes (1968) et ne présente pas un caractère foncièrement original. Les auteurs ont simplement rendu plus aisée l'application de la formule en substituant la consultation de la liste de 769 mots de Dale par le nombre de mots de plus de 6 lettres. Cette substitution est justifiée par une corrélation entre les deux variables de $0,78$, qui a été mesurée sur un corpus de 165 passages de 100 mots, tirés de dix livres de styles et de niveaux de difficulté variés.

Ils obtiennent la formule suivante (Barker et Stokes, 1968, 399) :

$$X_1 = 2,69 + 0,06\,X_2 + 9,55\,X_3 + 6,97\,X_4 \tag{2.15}$$

où les variables sont :

- X_1 **:** le niveau scolaire pour lequel est adapté le texte ;
- X_2 **:** le nombre moyen de mots par phrase ;
- X_3 **:** le ratio de syntagmes prépositionnels sur le nombre de mots ;
- X_4 **:** le ratio d'absents des 769 mots de la liste de Dale (1931)) par rapport au nombre total de mots. Il est toutefois estimé sous la forme du ratio de mots de plus de six lettres sur le nombre total de mots.

Cette substitution entraîne un gain de temps appréciable : pour un extrait de 100 mots, le temps de comptage passe de 102,9 secondes à 39,85 (Barker et Stokes, 1968, 400). Par contre, les auteurs ne sont pas précis sur la question du critère. On ne sait ni comment leurs passages ont été étalonnés, ni pourquoi le coefficient de X_4 et le terme indépendant sont aussi différents de la formule de Lorge (1944a).

L'année suivante, Tretiak (1969, cité par Klare, 1974, 68) propose une simplification plus radicale : laisser tomber la variable « nombre de syntagmes prépositionnels ». En effet, son dénombrement n'est pas suffisamment fiable et le facteur en lui-même apporte peu d'informations supplémentaires par rapport aux deux autres variables. Sur la base des leçons de McCall et Crabbs de 1961, il recalcule les coefficients de la formule de Lorge et obtient l'équation suivante :

$$X_1 = 10\,X_2 + 0,07\,X_3 \tag{2.16}$$

où les variables sont :

- X_1 **:** un niveau scolaire (même variable que l'équation 2.5) ;
- X_2 **:** le nombre moyen de mots par phrase ;

X_3 **:** le ratio d'absents des 769 mots de la liste de Dale (1931)) par rapport au nombre total de mots.

Les formules de Flesch Les formules de Flesch (1943 ;1948) vont connaître assez rapidement quelques révisions. En effet, certains chercheurs estiment encore possible de pousser plus loin la simplification opérée par Flesch. Ainsi, Farr *et al.* (1951) remplacent le comptage des syllabes par un décompte des mots monosyllabiques au sein de la RE. Cela permet de réduire le temps d'analyse et, surtout, évite que l'analyste ait à connaître les règles de syllabification. Dans ce cas aussi, la substitution est légitimée par un haut taux de corrélation ($r = 0,91$) entre ces deux facteurs. La formule recalculée à partir d'un échantillon de 360 passages de cent mots est la suivante (Farr *et al.*, 1951, 334) :

$$\text{New Reading Ease Index} = -31,517 + 1,599\, nosw - 1,015\, sl \tag{2.17}$$

où les variables sont :

$nosw$ **:** le nombre moyen de mots par phrase ;

sl **:** le nombre de mots monosyllabiques pour 100 mots.

La révision suivante s'appuie quant à elle sur un changement de critère, profitant de la réédition de 1950 des leçons de McCall et Crabbs. Ainsi, Powers *et al.* (1958) ont recalculé, afin de les moderniser et de les comparer sur la base d'un même critère, les quatre formules suivantes : l'indice de facilité de lecture de Flesch, les équations de Dale et Chall et de Farr-Jenkins-Paterson, ainsi que l'indice Fog de Gunning. Sur ce nouveau corpus, le coefficient de corrélation multiple de la formule de Flesch diminue légèrement, passant de $0,66$ à $0,635$; ceux des formules de Farr-Jenkins-Paterson et Gunning atteignent $0,58$; alors que celui de la formule de Dale and Chall augmente d'un centième par rapport à sa valeur initiale de $0,70$. Powers *et al.* (1958, 101) en concluent que c'est bien cette dernière formule qui est la plus précise, répliquant les résultats de Klare (1952).

La version modernisée de l'indice de facilité de lecture correspond à l'équation suivante (Powers *et al.*, 1958, 101) :

$$\text{Reading Ease} = -2,2029 + 0,0778\, X_2 + 0,0455\, X_3 \tag{2.18}$$

où les variables sont :

Reading Ease : le niveau scolaire moyen des élèves capables de répondre à 50% des questions de compréhension sur le texte ;

X_2 **:** le nombre moyen de mots par phrase ;

X_3 **:** le nombre de syllabes par 100 mots.

Cette nouvelle version sera critiquée par Harrisson (1980, 58-59), qui remarque qu'elle tend à plafonner et à ne prédire que très rarement un niveau de difficulté supérieur à la fin des primaires. La formule serait donc particulièrement adaptée pour les textes du primaire, mais discrimine mal les textes du secondaire.

Peu après, Coke et Rothkopf (1970) adaptent la RE de Flesch à une application automatisée. Comme le dénombrement des syllabes constitue à l'époque un problème algorithmiquement complexe, ils proposent de l'estimer via un comptage des voyelles, mais sans modifier autrement l'équation de Flesch.

La formule de Dale and Chall Il s'agit de la formule qui connaît le plus de révisions. Ce succès s'explique sans doute par la contradiction suivante : bien qu'étant considérée comme la plus fiable, cette formule est aussi plus fastidieuse à appliquer, puisqu'elle recourt à une liste de mots.

La première révision est due à Powers *et al.* (1958). Comme nous l'avons expliqué précédemment, ils vont moderniser la formule sur la nouvelle édition des leçons de McCall et Crabbs. Ils obtiennent les coefficients suivants (Powers *et al.*, 1958, 101) :

$$X_1 = 3,2672 + 0,0596\,X_2 + 0,1155\,X_3 \tag{2.19}$$

où les variables sont :

- X_1 **:** le niveau scolaire moyen des élèves capables de répondre à 50% des questions de compréhension sur le texte ;
- X_2 **:** le nombre moyen de mots par phrase ;
- X_3 **:** le pourcentage de mots absents de la liste de 3 000 mots de Dale.

Par la suite, Nyman *et al.* (1961) essaient de réduire la liste de Dale, afin de la rendre plus simple à utiliser. Pour ce faire, ils ne conservent que les mots qui sont également repris dans les 1 037 mots les plus fréquents de la liste de Thorndike-Lorge. Ils obtiennent ainsi 920 termes. Leur formule simplifiée – dont ils ne rapportent pas l'équation, mais qui est entraînée à partir de la version de 1950 des leçons de McCall et Crabbs – n'obtient qu'un coefficient de corrélation multiple de $0,59$, alors que la version originale atteint un R de $0,71$. Les auteurs en concluent que la perte de précision est trop importante et qu'il n'est pas souhaitable de réduire la liste de Dale, de cette façon du moins.

Cette problématique de la liste continue, par la suite, à faire couler de l'encre. Crane (1963, cité par Klare, 1974) étudie purement et simplement la possibilité de substituer la consultation de la liste par l'emploi d'une autre variable. Il expérimente à cette fin différents prédicteurs, dont le meilleur se révèle être « le nombre de mots de trois syllabes ou plus », une variable déjà employée par Gunning (1952). Il estime que cette nouvelle variable constitue

déjà un sérieux candidat de substitution. Toutefois, il observe que l'ajout de deux variables supplémentaires – le nombre de mots commençant par les consonnes *d*, *g*, *h*, *l*, *m*, *s*, *t* moins le nombre de mots commençant par *a*, *i* et *o* et le nombre de verbes abstraits – augmente encore l'efficacité de cette nouvelle formule, qui n'a toutefois plus grand chose de commun avec la formule de Dale et Chall.

Parallèlement, d'autres chercheurs, représentatifs de la volonté de l'époque de spécialiser les formules, s'évertuent à modifier la liste afin d'accommoder la formule originale à des types de textes particuliers. Ainsi, Brown (1965) estime que la formule surévalue la complexité des textes scientifiques, car la liste de Dale manque de termes scientifiques. Depuis que Dale a constitué sa liste, le vocabulaire scientifique a opéré une pénétration progressive dans la vie quotidienne, ce qui laisse penser que les élèves des années 60 connaissent davantage de termes scientifiques que leurs homologues des années 30. Ainsi, Brown (1965) montre que, en ajoutant seulement sept mots à la liste de Dale, on observe une diminution de deux niveaux scolaires pour des prédictions sur deux manuels de science.

Holmquist (1968, cité par Klare, 1974) pousse l'expérience plus loin. En se basant sur la nouvelle version des leçons de McCall et Crabbs (1961), il recalcule la formule de Dale-Chall et obtient un R de $0,69$. Ensuite, il dresse une liste de 102 termes scientifiques, connus de 80% des enfants de quatrième année primaire, dont il se sert pour établir la formule suivante, toujours sur la base du même critère (Klare, 1974, 72) :

$$X_1 = 3,4457 + 0,0510\,X_2 + 0,1195\,X_3 \tag{2.20}$$

où les variables sont :

X_1 **:** le niveau scolaire moyen des élèves capables de répondre à 50% des questions de compréhension sur le texte ;

X_2 **:** le nombre moyen de mots par phrase ;

X_3 **:** le pourcentage de mots absents de la liste révisée Dale, qui contient 3 102 termes.

Cette équation adaptée possède un R plus élevé d'un point que la formule originale recalculée par Holmquist. De plus, son erreur standard d'estimation[26] est également inférieure

26. L'erreur standard d'estimation est une statistique qui mesure l'erreur moyenne que l'on commet lorsqu'on utilise un modèle de régression linéaire pour effectuer une prédiction donnée. On peut aussi la décrire comme la variabilité qui subsiste après que l'on a utilisé les variables indépendantes, $X_1, \cdots, X_n$, pour prédire Y.

Elle se calcule selon la formule suivante (Howell, 2008, 251) :

$$s_{Y \cdot X} = \sqrt{\frac{\sum_i^N (Y_i - \hat{Y}_i)^2}{N-2}} \tag{2.21}$$

où le symbole $\hat{Y}_i$ représente la valeur prédite au moyen de la droite de régression pour l'observation *i*.

à celle de la formule classique. Holmquist en déduit que le passage du temps pourrait bien avoir détérioré les performances de l'indice de Dale et Chall et qu'il serait préférable de privilégier une version moderne de leur formule [27]. Notons que si Holmquist met à jour la possible vétusté d'un outil largement employé dans les milieux pédagogiques, on ne peut pas vraiment considérer qu'il propose une formule spécifique aux textes scientifiques, puisque son critère reste un ensemble de textes divers.

Enfin, Stocker (1971) suit le même raisonnement à propos d'un domaine radicalement différent : les textes religieux catholiques. Respectant la procédure selon laquelle la liste de Dale a été dressée, il identifie 204 mots connus de 80% des enfants de quatrième année primaire dans les écoles catholiques et les ajoute aux 3 000 mots de la liste de Dale.

L'indice Fog La dernière formule du quatuor classique, l'indice Fog, n'a pas connu le même succès auprès des réviseurs. Il faut dire qu'elle est déjà pensée pour être la plus simple possible à appliquer et qu'il semble difficile de pousser plus loin cette caractéristique. Dès lors, seuls Powers *et al.* (1958) vont proposer une nouvelle version du célèbre indice : ils le modernisent en l'entraînant sur la version de 1950 des leçons de McCall and Crabbs. La nouvelle équation est la suivante (Powers *et al.*, 1958, 101) :

$$\text{Fog Index} = 3,0680 + 0,0877\,X_1 + 0,0984\,X_2 \tag{2.22}$$

où les variables sont :

Fog Index : le niveau scolaire moyen des élèves capables de répondre à 50% des questions de compréhension sur le texte ;

X_1 **:** le nombre moyen de mots par phrase ;

X_2 **:** le pourcentage de mots polysyllabiques [28].

Les formules spécialisées

Parallèlement à ces révisions, dont certaines débouchent sur des formules spécialisées, apparaissent également de nouvelles formules qui se présentent comme spécialisées à un domaine particulier. On peut les regrouper selon les quatre domaines suivant : la lisibilité de tests standardisés ; les textes pour le début des primaires ; les textes à contenu scientifique et l'oral.

27. Cette hypothèse sera validé expérimentalement par la suite par Jacobson *et al.* (1978a).

28. L'article de Powers *et al.* (1958) n'est pas clair concernant le terme « polysyllabique ». Suivant le sens commun du terme, on aurait tendance à l'interpréter comme les mots comportant au moins **deux** syllabes. Cependant, étant donné que la formule de Gunning (1952) utilise comme prédicteur le nombre de mots de **trois** syllabes ou plus et que les auteurs disent réviser l'indice Fog sans en modifier le type des variables indépendantes, il semble qu'il faille plutôt entendre le terme « polysyllabique » dans cette acception un peu particulière.

La lisibilité des tests standardisés Johnson et Bond (1950) avaient mis en évidence, à l'aide de la formule de Flesch, le fait que divers tests standardisés, tels que la mesure du QI, contenaient fréquement des items dont le niveau de lecture dépassait la moyenne nationale. Ainsi, un sujet risquait d'échouer au test, non parce qu'il ne pouvait pas résoudre les problèmes qui lui étaient posés, mais parce qu'il ne les comprenait pas.

Dès lors, Forbes et Cottle (1953) décident de développer un outil qui permette de déterminer la lisibilité de divers tests standardisés. Pour ce faire, ils rassemblent un corpus de 27 tests communs à l'époque. Pour établir le niveau de difficulté de ce corpus, ils appliquent à chacun des tests cinq formules de lisibilité et leur attribue la moyenne des cinq prédictions. Cette façon de définir un critère apparaît évidemment criticable.

Comme prédicteur, Forbes et Cottle (1953) utilisent une mesure de fréquence des mots difficiles basée sur le *Thorndike Junior Century Dictionary*. Plus précisément, pour tous les mots qui ne sont pas compris parmi les 4 000 termes les plus fréquents, il s'agit de consulter ce dictionnaire où l'on trouve un indice de fréquence correspondant à la tranche de 1000 mots où se situe le lexème cible. Ainsi, un mot classé en 6 900^{e} position se voit attribuer un score de 7. Ensuite, on additionne ces valeurs avant de diviser la somme obtenue par le nombre total de mots dans l'échantillon. Le résultat est mis en correspondance, dans un tableau, avec le niveau scolaire nécessaire pour être capable de lire le test. Notons que seuls les aspects lexicaux sont considérés par ces deux auteurs.

Formules pour le début des primaires Une des conséquences de l'emploi quasi exclusif, à cette époque, des leçons de McCall and Crabbs est que les formules qui en découlent ne sont pas adaptées à la prédiction de matériaux pour le début des primaires. Il s'agit d'une tranche d'âge délaissée sur laquelle vont se pencher quelques chercheurs, dont le plus important est Spache (1953).

En 1953, George Spache conçoit une formule spécifique aux matériaux de la première à la troisième primaire. Il utilise comme critère un ensemble de 224 passages de 100 mots tirés de 152 livres utilisés communément au début des primaires. La difficulté de chaque extrait correspond au niveau scolaire auquel le livre est utilisé. L'équation qu'il obtient est la suivante (Spache, 1953, 412) :

$$X_1 = 0,839 + 0,141\, X_2 + 0,086\, X_3 \tag{2.23}$$

où les variables sont :

- X_1 **:** le niveau scolaire estimé du texte ;
- X_2 **:** le nombre moyen de mots par phrase ;
- X_3 **:** le nombre d'absents de la liste de 769 mots de Dale.

Elle se caractérise par un coefficient de corrélation multiple élevé, de $0,818$, qui assure un certain succès à cette formule, aussi bien aux États-Unis qu'en Angleterre (Harrisson, 1980, 67).

Néanmoins, Stone (1956) estime que la précision de la formule de Spache (1953) peut être améliorée en modifiant la liste de Dale. En effet, ayant réalisé une étude de fréquence sur un corpus de textes destinés aux premières années du primaire, il a constaté que de nombreux termes de cette liste sont trop complexes (ex. : *neither*, *company*, *fancy*...) et que d'autres (*pig*, *rabbit*, *doll*...), plus simples, en sont absents. Il propose dès lors une révision de la liste de Dale en substituant 173 mots difficiles par d'autres, plus simples. Utilisant cette liste modifiée avec la formule de Spache (1953), Stone (1956) obtient des prédictions inférieures pour les mêmes matériaux. Spache (1960) lui-même utilise cette nouvelle liste dans un ouvrage postérieur.

La même année, Tribe (1956, cité par Klare, 1963, 70) développe une autre formule pour les primaires, à partir des leçons de McCall et Crabbs de 1950. Il les utilise toutefois d'une façon originale : il administre effectivement une série de tests à 406 enfants, plutôt que de se reposer sur les valeurs données par les manuels, ainsi que l'ont fait ses prédécesseurs. Sur la base de ce corpus réévalué, il trouve un R multiple de $0,761$ pour l'équation suivante (Klare, 1963, 70) :

$$X_1 = 2,9347 + 0,0719\,X_2 + 0,1043\,X_3 \tag{2.24}$$

où les variables sont :

X_1 **:** le niveau scolaire estimé d'un élève capable de répondre correctement à 50% des questions à un test de compréhension portant sur ce texte ;

X_2 **:** le nombre moyen de mots par phrase ;

X_3 **:** le pourcentage de mots différents absents de la liste de Rinsland (*A Basic Vocabulary of Elementary School Children*).

Une troisième formule spécifique au début des primaires a été développée par Wheeler et Smith (1954). Les auteurs s'inspirent de Dolch (1948) en ce qui concerne le critère – le niveau de difficulté des passages correspond au niveau scolaire ciblé par le livre dont il est échantillonné – et de Johnson (1930), auquel ils reprennent l'idée d'évaluer la difficulté lexicale au moyen du nombre de mots polysyllabiques. Leur formule se présente ainsi :

$$X_1 = 10 * X_2 * X_3 \tag{2.25}$$

où les variables sont :

X_1 **:** un score abstrait qu'il faut interpréter en référence à une table qui donne le niveau scolaire correspondant ;

X_2 **:** le nombre moyen de mots par « unités » syntaxiques [29] ;

X_3 **:** le pourcentage de mots polysyllabiques.

Cette formule présente un aspect hautement original dans la technique statistique utilisée pour combiner les variables indépendantes. Il ne s'agit pas ici d'additionner leurs effets, mais de les multiplier, ce qui revient à considérer qu'il existe une interaction entre la longueur des « unités » et la longueur des mots. Cette technique ne sera plus réutilisée avant McLaughlin (1969), peut-être parce que sa pertinence n'est pas simple à établir, surtout lorsque seule l'interaction est prise en compte.

Formules pour des textes scientifiques À côté des efforts précités d'adaptation de la formule de Dale and Chall à la lisibilité de textes scientifiques, d'autres travaux se sont penchés sur la question. En 1961, Jacobson défend une thèse dans laquelle il présente des formules spécifiques à des manuels de chimie et de physique, qui visent plutôt la fin du secondaire et le début des études supérieures. Il rassemble son propre corpus constitué d'un ensemble d'échantillons issus de seize manuels de chimie et seize de physique. La difficulté de ces passages est établie en fonction du nombre de mots inconnus qu'ils contiennent, un test que Jacobson (1965, 450) appelle *Underlining Test*. Ces extraits servent ensuite à concevoir quatre formules qui utilisent diverses variables propres à évaluer la difficulté de textes de spécialité [30].

Une autre formule pour la difficulté à la lecture de matériaux scientifiques est due à Robert Shaw (1967, cité par Klare, 1974, 78). Cet auteur développe lui aussi une formule en se basant sur le nombre d'erreurs commises par des élèves de la quatrième à la sixième année primaire sur un corpus de matériaux scientifiques au programme de la quatrième. Avec l'aide de Jacobson, il implémente un programme apte à mesurer automatiquement 17 variables, parmi lesquelles trois se révèlent significatives suite à une procédure de sélection automatique de variable Shaw et Jacobson (1968).

Formules pour l'oral Pour en terminer avec ces formules spécialisées, citons rapidement les travaux de Rogers (1962) qui s'attachent à évaluer la difficulté de textes oraux. On y rencontre une discussion intéressante sur la meilleure unité syntaxique à considérer pour l'oral. La notion de phrase étant peu adaptée à ce type de discours, Rogers privilégie l'unité d'idée qu'il définit comme « une préposition indépendante, considérée avec ses modifieurs ». Cet aspect excepté, le reste de l'étude reste classique et la formule obtenue sur la base d'un

29. Sous le terme « unité », les auteurs entendent un ensemble de mots délimités par un point, un point d'interrogation ou d'exclamation, un point-virgule, deux points ou encore un tiret (Wheeler et Smith, 1954, 398).

30. Pour le détail de ces variables et des équations, qui sortent du cadre de cette étude, nous renvoyons le lecteur à Jacobson (1965, 455-456).

corpus oral est la suivante (Rogers, 1962, 220) :

$$X_1 = -2,0625 + 0,669\, I + 0,4981\, LD \tag{2.26}$$

où les variables sont :

X_1 **:** le niveau scolaire estimé d'un élève capable de répondre correctement à 50% des questions à un test de compréhension portant sur ce texte ;

I **:** le nombre moyen de mots par unité d'idée ;

LD **:** le nombre de mots absents de la liste de Dale de 3 000 mots.

Quelques années plus tard, Fang, qui semble ignorer les travaux de son prédécesseur, défend une thèse sur le même sujet. Il conçoit l'Easy Listening Formula (ELF), qui correspond simplement à la moyenne de la somme des syllabes par phrases qui appartiennent à des mots de plus d'une syllabe, la première syllabe n'étant pas prise en compte. La valeur moyenne à ne pas dépasser pour être compris de la majorité des auditeurs est de 12 (Fang, 1966, 68).

Nouvelles formules généralistes

Enfin, nous envisageons dans cette dernière sous-section les quelques formules qui adoptent une visée généraliste. Elles sont particulièrement intéressantes, car elles illustrent deux tendances qui se développeront par la suite : le désir d'automatisation et l'emploi de graphes à la place d'une équation.

La première d'entre elles, due à Bloomer (1959) constitue une nouvelle tentative de s'attaquer à la question de l'abstraction. Elle s'inspire de Boder (1940), qui a montré que le ratio des adjectifs par rapport aux verbes augmente à mesure que le degré d'abstraction d'un texte croît. Pour Bloomer, le niveau d'abstraction serait lié à la quantité de modifieurs [31]. Dès lors, il étudie les corrélations entre la difficulté de textes issus de manuels scolaires de la première à la sixième primaire et les quatres variables suivantes :

– Le nombre de mots par modifieur (plus il est réduit, plus le texte est supposé être abstrait) ;
– L'indice de complexité sonore, basé sur la similarité des sons entre les lettres adjacentes ;
– L'indice de complexité formelle, basé sur la similarité des formes entre les lettres adjacentes [32] ;
– Le nombre de lettres par mot.

Il conclut que les meilleurs prédicteurs sur son corpus sont l'indice de complexité sonore et le nombre de mots par modifieur. En les combinant, il obtient un R de $0,78$, un résultat qu'il juge similaire, en magnitude, à ceux de Flesch ou Lorge (Bloomer, 1959, 269).

31. Sous le terme « modifieur », il englobe la plupart des adjectifs et des adverbes. Pour le détail des exceptions, se reporter à Bloomer (1959, 269).

32. Pour une description plus détaillée de ces deux indices, se référer à Bloomer (1956).

Dans la lignée de Flesch et Dale and Chall, Smith (1961) conçoit une formule destinée aux élèves de la deuxième primaire à la deuxième secondaire qu'il appelle l'index Devereaux, d'après le nom de la fondation où il travaille. Toujours dans une optique de simplification, il défend l'idée que le nombre de lettres par mots représente un substitut valable à la consultation d'une liste ou au comptage des syllabes. Non seulement, le dénombrement de cette variable ne requiert aucune connaissance linguistique, mais surtout, il est aisément automatisable. Ce souci de n'utiliser que des variables dont la technique de paramétrisation est parfaitement explicitable fait de Smith le premier représentant de cette tendance à l'automatisation, laquelle revêt une importance particulière dans le cadre de cette thèse.

La formule, entraînée sur un échantillon de passages extraits de 17 livres pour enfants, est la suivante (Smith, 1961, 299) :

$$X_1 = -6,49 + 1,56\,WL + 0,19\,SL \tag{2.27}$$

où les variables sont :

X_1 **:** le niveau scolaire estimé d'un texte ;

WL **:** le nombre moyen de lettres par mots [33] ;

SL **:** le nombre de mots par phrase.

Il obtient un R de $0,74$, ce qui représente un résultat tout à fait correct par rapport à la littérature de l'époque.

La formule de Smith (1961) sera reprise et implémentée par Danielson et Bryan (1963) sur un ordinateur UNIVAC 1105, lequel utilisait des cartes perforées. Plus précisément, les deux auteurs ont d'abord cherché à automatiser la formule de Farr *et al.* (1951) en implantant des routines de détection des mots monosyllabiques et des phrases. Leur démarche dénote une remarquable réflexion linguistique sur les moyens de prendre en compte les exceptions qui posent aujourd'hui encore problème en TAL. Toutefois, les auteurs se sont vite rendu compte que les capacités des machines de l'époque rendaient la chose malaisée et ils se sont rabattus sur des variables plus simples à automatiser et fondées sur le comptage des caractères.

En prenant comme critère les leçons de McCall and Crabbs (1950), ils ont établi une variante de la formule de Smith (1961), dont l'équation est la suivante (Danielson et Bryan, 1963, 206) :

$$X_1 = -0,6059 + 1,0364\,CPSp + 0,0194\,CPSt \tag{2.28}$$

où les variables sont :

33. À ce propos, notons que Smith (1961, 299) compte les mots en fonction du nombre d'espaces dans la phrase.

X_1 **:** le niveau scolaire moyen des élèves capables de répondre à 50% des questions de compréhension sur le texte ;

$CPSp$ **:** le nombre moyen de caractères par espace ;

$CPSt$ **:** le nombre d'espaces par phrase.

Cette formule, où la nature du comptage fait apparaître l'outil comme une version plus fruste de ses prédécesseurs, obtient tout de même un R de $0,575$. Il s'agit certes d'un résultat nettement inférieur à ceux obtenus par Dale and Chall ou Flesch, mais comparable à la formule de Farr *et al.* (1951), laquelle nécessite pourtant des opérations algorithmiques bien plus complexes. Ainsi, Danielson et Bryan (1963) ont été parmi les premiers à démontrer que l'automatisation des formules de lisibilité était possible. Ils estiment que grâce à l'économie de temps réalisée pour l'analyse de textes, le recours à une machine pourrait permettre aux concepteurs de nouvelles formules d'accroître la taille de leur corpus d'entraînement et, par voie de conséquence, l'exactitude de leurs résultats.

Parallèlement à cette automatisation, Fry (1963) innove dans la recherche de la simplicité en publiant, dans son livre *Teaching Faster Reading : A Manual*, un graphique qui permet de se dispenser des calculs que requièrent les formules classiques. Il suffit de croiser deux facteurs classiques – le nombre de phrases pour 100 mots et le nombre de syllabes pour 100 mots – sur ce graphe pour situer le niveau de difficulté par rapport à une année scolaire ou, et c'est un autre intérêt de l'outil, par rapport à un autre livre. Par la suite, Fry (1968) publie une révision de sa formule afin de l'adapter aux *grades* américains. Il montre également que les résultats de sa formule sont fortement corrélés avec ceux d'autres formules bien connues.

À côté de ces deux tendances, Botel (cité par Schneyer, 1963) propose une approche lexicale qui vise également la simplicité. Après avoir construit une liste de 1185 mots usuels à partir d'un corpus de livres destinés aux enfants de la première à la troisième primaire, il recherche un moyen de l'utiliser pour évaluer la difficulté de textes. Il observe que, pour chaque échantillon qu'il analyse, la distribution du vocabulaire par *grade* présente une caractéristique commune : un échantillon donné comporte environ 9% de mots appartenant à des niveaux scolaires supérieurs à celui qui a été assigné au livre dont il est extrait. Botel en conclu que la difficulté d'un texte donné correspond au niveau médian des échantillons prélevés sur ce texte. Au final, cette technique n'est guère différente de l'approche qui consiste à considérer le nombre de termes absents d'une liste de référence. D'autant que, pour les niveaux supérieurs à la troisième année et donc, non couvertes par sa liste, Botel propose de convertir le nombre de termes absents en un niveau scolaire à l'aide d'une table.

Conclusion

À la lecture des travaux présentés dans cette section, notre impression est que le domaine manque quelque peu de souffle au cours de cette troisième période. Bormuth (1966, 124) va même jusqu'à dire que les recherches majeures se sont arrêtées en 1948. D'une part, la majorité des recherches se contentent d'optimiser les formules de la période classique ou de les spécialiser. D'autre part, les fondements théoriques de l'approche classique ne sont guère remis en cause. Ce n'est d'ailleurs pas un hasard si très peu de formules qu'elle a vu naître sont restées célèbres, à l'exception de celles de Spache (1953), Powers *et al.* (1958) et Fry (1963).

Cependant, en s'arrêtant à certains détails et en les remettant en perspective, on se rend compte que le bilan n'est pas si négatif. On perçoit alors cette troisième période comme une étape de transition, où s'opère à la fois la digestion les formules classiques – leurs forces et faiblesses apparaissent plus clairement, notamment grâce à des études de validation dont la plus représentative est due à Klare (1952) –, et la gestation de la période suivante. Ainsi, les chercheurs se libèrent de la domination des leçons de McCall et Crabbs et s'interrogent à nouveau sur d'autres formes de critère, ce qui va conduire à ce que nous appelons la révolution de la closure. Quelques travaux explorent également de nouvelles perspectives statistiques, dont la prise en compte de l'interaction (Wheeler et Smith, 1954). Enfin, les efforts d'automatisation de Danielson et Bryan (1963) préparent une seconde révolution, technologique celle-là, qui va modifier les pratiques de recherche dans le domaine.

Toutefois, avant d'aborder cette quatrième période, que nous avons intitulée *la révolution de la closure*, il est nécessaire de s'arrêter sur les premiers travaux parus pour le français, car ils ressortissent plutôt à cette période de transition.

L'arrivée de la lisibilité dans le monde francophone

Les formules que nous avons présentées jusqu'ici ont été essentiellement développées aux États-Unis. Il faut attendre l'année 1956 pour qu'André Conquet fasse connaître aux Français, dans un ouvrage intitulé *Lisibilité*, le domaine de la lisibilité au travers des formules de Flesch (1948) et de Dale et Chall (1948a). Deux ans plus tard, Kandel et Moles (1958) mettent au point ce qui est communément considéré comme la première formule de lisibilité pour le français.

Il s'agit en réalité d'une simple adaptation de la formule de Flesch, appliquée au contexte particulier de la compréhension de textes radiophoniques, une optique qui rappelle celle de Fang (1966). Cette adaptation s'appuie sur un corpus de onze livres anglais ou français ac-

compagnés de leurs traductions respectives, ce qui permet de comparer la difficulté d'un même texte entre les deux langues. Kandel et Moles (1958) observent que, dans les textes francophones, les mots sont en moyenne $1,15$ fois plus longs que dans les textes anglophones. Ils ajoutent :

> *Quant à la longueur moyenne des phrases, les différences paraissent beaucoup moins systématiques et certainement beaucoup moins reliées à des variables proprement linguistiques qu'au style propre de tel ou tel traducteur (Kandel et Moles, 1958, 257)* .

Dès lors, ils proposent la version suivante de la formule de Flesch :

$$X_1 = 207 - 1,015\, lp - 0,736\, lm \tag{2.29}$$

où les variables sont :

X_1 **:** un score compris entre 0 et 100 (voir formule 2.9) ;

lp **:** le nombre moyen de mots par phrase ;

lm **:** le nombre de syllabes pour 100 mots.

Il s'agit, ni plus ni moins, de la RE de Flesch (1948) où seul le coefficient de lm a été divisé par $1,15$. Les auteurs sont conscients des limites de leur approche : la taille de l'échantillon utilisé est réduite ; il est nécessaire de postuler la fidélité de la traduction quant au niveau de difficulté, ce qui constitue déjà une hypothèse forte, mais il faut également considérer que le public francophone correspond à celui qui a été modélisé par Flesch (1948), ce qui n'est visiblement pas le cas. Enfin, il n'est pas certain que l'importance relative des deux variables soit la même dans les deux langues. « Il se pourrait, par exemple, que la longueur des phrases soit en français un facteur de lisibilité plus important que la longueur des mots » (Kandel et Moles, 1958, 258).

La recherche de Kandel et Moles (1958), parue dans une publication confidentielle, n'a guère eu d'écho dans l'enseignement ou l'édition. C'est Gilbert de Landsheere (1963) de l'université de Liège qui va réellement diffuser la problématique de la lisibilité dans le monde francophone, sans toutefois révolutionner le domaine. Il reprend telle quelle l'équation de Flesch (1948) et son échelle de mesure graduée de 0 à 100 et modifie quelque peu les règles de comptage : toutes les syllabes sont prises en compte, y compris celles qui comprennent un *e* muet, tandis que la définition de phrase recouvre toute unité de pensée délimitée par un point, un point d'exclamation ou un point d'interrogation. Gilbert de Landsheere (1963) présente également l'indice d'intérêt humain de Flesch, auquel il n'apporte guère de modifications.

On voit donc que cette première phase des recherches sur le français est également carac-

térisée par un faible degré d'innovation. On se contente d'adapter la formule de Flesch en modifiant quelques détails, mais sans chercher à recalculer les coefficients à partir d'une population de lecteurs francophones. Ce qui n'empêche pas certains pédagogues d'utiliser ces premières formules pour étalonner des matériaux d'entraînement à la lecture. Citons pour exemple le fichier ATEL de Foucambert et André (détails *cf.* Benoît, 1986b, 52).

Toutefois, se restreindre à une simple adaptation de Flesch ne constitue pas une position tenable pour les chercheurs en lisibilité qui s'intéressent au français. Aussi bien Kandel et Moles (1958) que de Landsheere (1963) en sont conscients, ce qui va pousser ce dernier à poursuivre les recherches. Il sera aidé d'un doctorant, George Henry, qui deviendra le concepteur de la première formule de lisibilité spécifique au français. Nous repoussons toutefois l'exposition de ses travaux à la fin de la prochaine section, car la méthodologie utilisée est caractéristique des changements qui vont s'opérer au cours de la période suivante, celle qui voit l'ascension de l'informatisation et du test de closure.

2.2.5 La révolution de la closure (1965 à 1979)

Introduction

L'année 1965 est généralement considérée comme un tournant dans l'évolution de la lisibilité, car elle marque l'avènement d'un nouveau type de critère : le test de closure. Il s'agit toujours d'un moyen de mesurer le niveau de compréhension d'une population pour un texte donné, mais qui est plus simple à concevoir, plus fiable et dont la correction est moins sujette à l'interprétation. Nous en résumons les grandes caractéristiques ci-après et discutons son application en lisibilité au Chapitre 4.

On peut parler de révolution de la closure, car ce nouveau test est employé dans de nombreuses études (Aquino *et al.*, 1969; Bormuth, 1966, 1969a; Carver, 1974a; Caylor *et al.*, 1973; Kincaid *et al.*, 1975) et produit des indices de validité élevés. Bien sûr, les autres types de critères rencontrent encore un certain succès, en particulier le jugement humain (Björnsson, 1983; Elley, 1969; Harris et Jacobson, 1973; Smith et Senter, 1967), mais aussi les tests de compréhension (Harris et Jacobson, 1976) ou les leçons de McCall and Crabbs (McLaughlin, 1969) [34].

Simultanément, une seconde révolution, technologique celle-là, s'opère suite aux progrès réalisés en informatique. Les formules de l'époque sont caractérisées par une automatisation croissante, aussi bien lors des étapes de conception que lors de leur application. L'ordinateur permet de s'affranchir de certaines limites computationnelles pour les analyses statistiques,

34. Se reporter au Chapitre 4 pour une description détaillée et une discussion de ces différents critères.

mais aussi de pallier le manque de fiabilité des comptages manuels. Dès lors, plutôt que d'optimiser une formule à deux ou trois variables en fonction de l'efficacité et de la simplicité, quelques chercheurs exploitent les possibilités de l'informatique pour définir un assortiment de formules. Celles-ci sont déclinées tantôt en fonction du niveau scolaire, tantôt en fonction du nombre de variables explicatives incluses.

On assiste ainsi à la réapparition de formules comportant plus de deux – à la rigueur trois – variables indépendantes (Bormuth, 1969a; Harris et Jacobson, 1976; Liau *et al.*, 1976), suite à la remise en cause de la conviction qu'il ne sert à rien d'en combiner davantage. Au contraire, Bormuth (1966, 1969a) défend l'idée que le nombre de facteurs dans la plupart des formules est trop réduit.

Notons enfin que ces nouvelles tendances coexistent avec des approches plus traditionnelles, notamment parce que l'ordinateur est loin d'être répandu chez les enseignants, pédagogues ou éditeurs. Ainsi, certains concepteurs de formules continuent de se focaliser sur la simplicité, qu'ils obtiennent au moyen de graphes (Fry, 1977; Mugford, 1970), de simplifications calculatoires (Anderson, 1983; Björnsson, 1983; Caylor *et al.*, 1973; McLaughlin, 1969), ou à l'aide de nouvelles variables, supposées plus prédictives (Aquino *et al.*, 1969; Botel et Granowsky, 1972; Elley, 1969; Jacobson, 1974).

Le test de closure

Des années vingt au début des années cinquante, différents critères ont été employés dans le développement des formules de lisibilité. Le plus courant d'entre eux était le test de compréhension basé sur des questions ouvertes ou à choix multiples. Toutefois, des voix se sont élevées pour faire remarquer que les difficultés éprouvées par un étudiant confronté à ce type de test peuvent tout autant être attribuées à la formulation des questions qu'à la complexité du texte.

En réponse à ce débat, Taylor (1953) présente un nouvel outil pour évaluer la compréhension d'un texte : le test de closure. Il consiste à retrouver une série de mots qui ont été éliminés de manière systématique dans un texte. Voici un exemple de test de closure portant sur un extrait de *Madame Bovary* de Flaubert, dans lequel un mot sur cinq a été supprimé :

> Ils arrivèrent, en effet, ... fameux Comices ! Dès le ..... de la solennité, tous ..... habitants, sur leurs portes, ... des préparatifs ; on avait ..... de lierres le fronton ..... la mairie ; une tente ..... un pré était dressée ..... le festin, et, au ..... de la place, devant ..... église, une espèce de ..... devait signaler l'arrivée ..... M. le préfet et ..... nom des cultivateurs lauréats.

Taylor (1953) nomme ce nouvel instrument en référence à l'idée de clôture, issue de la psychologie de la forme (Gestalt). Dans le cadre de cette théorie, l'individu « organise son champ perceptif selon cinq lois, dont la loi de closure, une tendance à compléter mentalement une forme familière incomplète » (Cornaire, 1985). Cette filiation a toutefois été critiquée par Weaver et Kingston (1963). Ceux-ci rappellent que la notion de closure est avant tout reliée aux perceptions du sujet et désigne donc un processus inconscient, ce qui n'est pas le cas du processus mental mis en œuvre lors d'un test de closure.

Malgré ces considérations sur son appellation, le test de closure est utilisé dès la fin des années cinquante pour mesurer la compréhension de textes. Pour Taylor (1953, 417), il constitue un sérieux concurrent des formules de lisibilité, puisqu'il peut prendre en compte une série de dimensions du processus de lecture que les formules ignorent : les connaissances préalables du lecteur, sa motivation, mais aussi le manque de cohérence du texte, ses imperfections stylistiques, etc.

Reste à déterminer si le test de closure mesure effectivement la compréhension à la lecture, les avis étant partagés à ce sujet. Weaver et Kingston (1963) concluent que « les tests de closure ne sont que peu à modérément corrélés à des tests de compréhension classiques » et que c'est plutôt le niveau de redondance textuelle qui est mesuré. Par la suite, Bormuth (1969b, 365) estime pour sa part :

> *It appears [...] that cloze tests made by deleting every fifth word measure skills closely related or identical to those measured by conventional multiple-choice reading comprehension tests. Little or no evidence was found to suggest that the two types of tests might measure different skills.*

Au final, l'idée que le test de closure constitue une mesure fiable de la compréhension à la lecture semble avoir été généralement acceptée, comme l'indiquent les études rapportées par Henry (1975, 55). La plupart d'entre elles obtiennent des corrélations nettement significatives entre les résultats à des tests de closure et des tests de compréhension classiques.

Par contre, le test de closure présente un défaut majeur lorsqu'il est employé pour évaluer la lisibilité de textes. Il est nécessaire de disposer d'un nombre suffisant de sujets, représentatifs de la population de lecteurs ciblés. En référence à la distinction opérée en début de chapitre, on peut dire que le test de closure permet de **mesurer** la lisibilité d'un texte, mais pas de la **prédire**. Dès lors, on comprend pourquoi le développement de formules de lisibilité n'a pas périclité suite à l'apparition de ce nouveau procédé.

En fait, le test de closure a rapidement été perçu comme un dispositif capable d'améliorer la qualité des critères utilisés pour concevoir de nouvelles formules. C'est Coleman (1965, cité par Klare, 1974, 77), Miller et Coleman (1967) et surtout Bormuth (1969a), qui vont

poser les bases méthodologiques en ce sens. Ainsi, Bormuth (1966) montre que le test de closure, en plus de favoriser la conception de formules plus fiables, permet également de ne plus limiter l'analyse de la difficulté à des passages, mais de s'attacher à des unités plus petites, telles que la phrase ou même le mot.

Ces auteurs discutent également d'un autre problème découlant de l'emploi de tests de closure comme critère, à savoir la question de l'interprétation des résultats. En effet, les formules développées sur la base de ce type de critère se servent du pourcentage de bonnes réponses moyennes obtenues par leurs sujets sur un texte donné. Or, cette échelle manque de pragmatisme pour les enseignants, plus habitués à manipuler des niveaux scolaires. C'est pourquoi plusieurs chercheurs se penchent sur le moyen d'exprimer ces pourcentages en années scolaires, allant jusqu'à proposer des formules de conversion.

Coleman : les premières formules basées sur le test de closure

En 1965, Coleman publie quatre formules de lisibilité. Elles sont surtout remarquables parce qu'elles sont les premières à avoir été mises au point à l'aide du test de closure. Elles sont également caractéristiques de l'informatisation croissante du domaine, puisque le nombre des variables indépendantes retenues augmente et que leur paramétrisation est simple à automatiser. Les quatre formules prennent la forme suivante (Szalay, 1965, 965) [35] :

$$\begin{aligned} C\% &= -38,45 + 1,29\,w \\ C\% &= -37,95 + 1,16\,w + 1,48\,s \\ C\% &= -34,02 + 1,07\,w + 1,18\,s + 0,76\,p \\ C\% &= -26,01 + 1,04\,w + 1,06\,s + 0,56\,p - 0,36\,prep \end{aligned} \tag{2.30}$$

où les variables sont :

- $C\%$ **:** le pourcentage de lacunes correctement complétées lors d'un test de closure appliqué au texte ;
- w **:** le nombre de mots d'une syllabe dans un échantillon de 100 mots ;
- s **:** le nombre de phrases pour 100 mots ;
- p **:** le nombre de pronoms pour 100 mots ;
- $prep$ **:** le nombre de prépositions pour 100 mots.

Coleman (1965, cité par Klare, 1974, 78) rapporte, sur son échantillon d'entraînement, des R élevés pour ces quatres formules : respectivement $0,86$, $0,89$, $0,90$ et $0,91$. Suspicieux face

35. N'ayant pas eu accès directement au travail de Coleman (1965), nous reprenons ces équations à Szalay (1965).

de tels résultats, Szalay (1965) effectue une validation sur sept nouveaux passages et obtient les valeurs suivantes : $0,83$, $0,88$, $0,87$ et $0,89$. Il n'y a donc pas de réduction notable, ce qui montre que le recours aux tests de closure comme critère permettrait d'atteindre des coefficients de corrélation multiples plus élevés que l'emploi des leçons de McCall et Crabbs, par exemple.

Bormuth ou l'esprit de synthèse

Bormuth est sans doute le chercheur qui s'est attaqué le plus profondément à la question de la lisibilité. Ses études, dont Klare (1974, 80) estime qu'elles sont les plus exhaustives jamais publiées, présentent non seulement un grand intérêt en tant que telles, mais ont également joué un rôle revitalisant, posant de nouvelles fondations empiriques pour la discipline.

Dans un premier article, Bormuth (1966) s'attaque à une série de problèmes méthodologiques essentiels en lisibilité. Pour réaliser cette investigation, il collecte une série d'extraits et les étalonne grâce au test de closure selon la procédure suivante. Tout d'abord, il applique la formule de Dale et Chall (1948a) à divers passages de 275 à 300 mots afin d'en retenir quatre pour chaque niveau entre le quatrième et le huitième *grade*. De la sorte, il s'assure d'une représentation équilibrée de la difficulté au sein du corpus. Ensuite, cinq tests de closure différents – la suite de mots supprimés étant décalée d'un mot à chaque fois – sont préparés pour chacun des vingt passages retenus. Chacun d'eux est alors administré à 675 étudiants de la quatrième primaire à la deuxième secondaire (ce qui fait 135 sujets par *grade*) supposés représentatifs de la population scolaire américaine et dont le niveau de compétence à la lecture varie du deuxième au douzième *grade*. Bormuth obtient ainsi vingt extraits étalonnés, et dont la complexité de chaque mot et de chaque phrase est également connue grâce au test de closure.

Sur la base de ce critère, Bormuth (1966) effectue plusieurs observations cruciales. En résumé, on peut dire qu'il obtient plusieurs résultats qui élargissent considérablement les perspectives de conception d'une formule de lisibilité. Ainsi, alors que ses prédécesseurs avaient toujours estimé que la relation entre la mesure de la compréhension et les différents prédicteurs est linéaire, Bormuth (1966, 98-102) juge que celle-ci est plutôt curvilinéaire, du moins au niveau des mots [36]. Deuxièmement, il observe que l'importance des variables ne varie pas en fonction du niveau de compétence des utilisateurs, ce qui suggère qu'une seule formule pourrait être utilisée pour un public, quelle que soit leur aptitude à la lecture.

Dans la même étude, Bormuth (1966) explore un ensemble de nouvelles variables (*cf.* Cha-

36. Notons que ce fait avait déjà été pressenti par Dale et Chall (1948a), qui avaient fourni une table de correction pour les résultats de leur formule.

pitre 3) et s'attache à la conception de formules de lisibilité pour des unités telles que la phrase ou même le mot, ce qui n'avait jamais été réalisé jusque-là. Il obtient respectivement des R de $0,665$ et $0,505$, ce qu'il considère comme suffisant pour défendre la possibilité de ce niveau de prédiction. Quant au niveau des extraits, il obtient le plus haut taux de corrélation multiple à cette date : $0,934$. Sans compter qu'il estime encore possible d'améliorer cette performance. Suite à ses expérimentations, il en vient à une nouvelle conclusion qui est le contre-pied de la méthodologie classique (Bormuth, 1966, 113) :

> *It is apparent that available readability formulas contain too few variables and thereby ignore variables of prose style which can make important independent contributions to the multiple correlations.*

Le projet suivant de Bormuth (1969a) consiste à étudier plus spécifiquement divers indices de lisibilité, afin d'identifier parmi eux les causes réelles la complexité textuelle, mais aussi de proposer des formules qui profitent de ses remarques méthodologiques précédentes. Nettement plus ambitieux, Bormuth collecte rien moins que 330 passages d'environ 100 mots, traitant de dix disciplines différentes, et destinés à des étudiants dont le niveau varie de la première primaire aux débuts de l'enseignement supérieur. Chacun des passages est transformé en un test de closure, décliné selon les cinq versions possibles, ce qui donne 1650 tests différents. Ceux-ci sont ensuite administrés à 2600 sujets de la quatrième primaire à la sixième secondaire dont les capacités à la lecture ont été préalablement évaluées au moyen du *California Reading Achievement Test*.

Sur la base d'un tel corpus d'entraînement, il examine l'efficacité de 169 variables, dont une majorité sont nouvelles et basées sur des théories psycholinguistiques. Il détermine pas moins de 24 formules de lisibilité[37] organisées selon trois axes : la taille de l'unité linguistique analysée (passage, phrase, unité minimale de ponctuation ou mot), le contexte d'utilisation et le taux de compréhension minimum. Le terme « contexte » désigne ici la manière dont la formule va être appliquée : soit par un scientifique patient et consciencieux, soit par un ordinateur, soit manuellement, par un analyste qui ne maîtrise pas les typologies linguistiques utilisées par Bormuth. Le dernier axe ressortit quant à lui d'une question fondamentale que pose Bormuth (1969a) : quel est le résultat à obtenir à un test de closure pour faire preuve de la bonne compréhension d'un passage donné ? À la suite de trois expériences, il obtient pour ce seuil les trois valeurs suivantes : 35%, 45% et 55%, dont il se sert pour exprimer le résultat de ses équations sous la forme de niveaux scolaires.

Au-delà de la tâche monumentale effectuée par Bormuth (1969a) rien qu'au niveau de la quantité (taille du corpus, nombre de variables et de formules), ses travaux se distinguent

37. Rapporter ici les équations calculées par Bormuth (1969a, 149-158) demanderait trop d'espace. Par conséquent, nous invitons le lecteur intéressé à se référer directement au rapport en question.

par diverses avancées méthodologiques dont certaines s'appuient sur les réflexions de 1966. Ainsi, il est le premier à expérimenter d'autres relations que la simple linéarité entre le critère et les prédicteurs. On rencontre ici des régressions polynomiales, où certaines variables sont élevées au carré ou au cube. Bormuth est aussi le premier dans l'histoire de la lisibilité à rapporter des taux de corrélations estimés sur un échantillon de test et non directement sur le corpus de développement. Ce faisant, il en vient à raffiner sa position de 1966 sur le nombre de variables à insérer dans une formule. S'il reste convaincu de la nécessité de sélectionner plus de deux prédicteurs, il observe que des formules trop complexes (qui comptent de 14 à 20 variables) souffrent de sur-apprentissage, car leur R chute drastiquement en validation croisée [38]. Notons que les taux de corrélation multiple qu'il rapporte sont les plus élevés qui aient été atteints jusque-là et qu'ils oscillent entre $0,928$ et $0,823$ pour le niveau des extraits. Aussi, au vu de la finesse de ses réflexions théoriques et de la richesse de ses expériences, il n'est pas étonnant qu'il soit considéré comme l'un des chercheurs majeurs en lisibilité.

Quelques travaux de moindre importance

Durant cette seconde moitié des années soixante, apparaissent quelques travaux de moindre rayonnement. Baker, Bloomer and Schwimmer (non daté, cités par Klare, 1974, 78-79) développent READSCORE, l'une des premières formules automatisées. On y découvre, comme chez Coleman, une augmentation du nombre de variables retenues. Cette fois, l'équation de régression (en réalité, seuls les coefficients sont rapportés) comprend cinq variables, dont l'une se distingue par sa nouveauté. Il s'agit de l'écart-type moyen de la longueur des mots du texte. En effet, en accord avec les travaux de Dolch (1948), Baker et ses collaborateurs observent que la moyenne seule ne suffit pas à discriminer correctement entre certains textes dont la longueur moyenne des mots est identique, mais répartie différemment. Ils emploient ainsi ces deux indicateurs statistiques au sein de leur formule, laquelle obtient un taux de corrélation multiple de $0,72$ sur leur corpus. Celui-ci est constitué de 130 passages issus des leçons de McCall et Crabbs.

Plus connu sous le nom de Automated Readability Index (ARI), l'indice de lisibilité développé par Smith et Senter (1967) pour l'armée américaine se distingue des autres approches automatisées de l'époque. Les deux auteurs équipent une machine à écrire électrique d'un mécanisme capable de compter le nombre de touches enfoncées, le nombre de mots (au moyen de la barre d'espace) et le nombre de phrases, la fin d'une phrase devant toutefois être indiquée manuellement à l'aide du signe égal. Par ailleurs, ils étalonnent leur corpus d'entraînement, constitué de vingt-quatre livres, selon un critère rarement utilisé en lisibi-

38. Il s'agit d'une technique d'évaluation de la capacité de généralisation d'un modèle. Elle consiste à appliquer celui-ci sur des données différentes de celles ayant servi à l'entraîner, ce qui permet d'obtenir des résultats sans biais.

lité : l'avis d'éditeurs. La formule qui en découle est la suivante :

$$X_1 = -21,43 + 0,5\,X_2 + 4,71\,X_3 \tag{2.31}$$

où les variables sont :

X_1 **:** le niveau scolaire que des éditeurs spécialisés attribueraient à ce texte ;

X_2 **:** le nombre moyen de mots par phrase ;

X_3 **:** le nombre moyen de caractères (y compris la ponctuation et les nombres) par mots.

En 1968, Milagros Aquino publie une étude qui expose sept formules de lisibilité obtenues en combinant diversement huit variables linguistiques. Parmi ces dernières, il suggère, comme nouvel indice de la difficulté, la proportion de suffixes d'origine latine. Celle-ci se révèle le meilleur prédicteur des performances d'étudiants de l'enseignement supérieur à un test de closure parmi les cinquante-neuf variables expérimentées (Aquino *et al.*, 1969, 1). Ainsi, la formule suivante, qui ne recourt qu'à la proportion de suffixes latins, est corrélée à $0,91$ avec le critère (Klare, 1974, 82) :

$$X_1 = 66,857 - 1,348\,X2 \tag{2.32}$$

où :

X_1 **:** le pourcentage de lacunes qui seraient correctement complétées à un test de closure portant sur ce texte ;

X_2 **:** la proportion de suffixes latins.

La même année, Björnsson fait paraître en suédois un article où il présente une nouvelle formule de lisibilité qu'il nomme l'indice LIX. Ces initiales sont simplement l'abréviation de Läsbarhetsindex, ce qui signifie « indice de lisibilité » en suédois. Sa formule est très simple ; nous la reprenons d'un article postérieur traduit en anglais (Björnsson, 1983, 481) :

$$LIX = \text{longueur moyenne des phrases} + \text{longueur des mots} \tag{2.33}$$

La longueur des mots est ici calculée comme le pourcentage de mots de plus de six lettres. Cette formule très simple émane d'une méthodologie atypique, mais stricte : une centaine de passages à la difficulté très variée ont été sélectionnés dans divers types de littérature. Leur difficulté a ensuite été estimée par une trentaine de sujets au moyen d'une échelle comportant neuf degrés, avant que ne soient étudiées les corrélations entre ces estimations et les deux variables précitées.

Björnsson (1983, 483) effectue une observation très intéressante. Ayant scindé en deux son corpus, il calcule trois équations de régression : une pour chaque moitié du corpus et une

pour l'ensemble des textes. Il obtient des coefficients sensiblement différents et en déduit que la régression est fortement dépendante de la composition du critère et n'est pas adéquate pour la conception d'une formule de lisibilité. Dès lors, il accorde la même importance aux deux facteurs de sa formule et obtient néanmoins un R de $0,92$ par rapport au critère.

Par la suite, la formule va être appliquée à diverses langues – français, anglais, allemand, danois, finnois –, toujours en respectant cette même méthodologie. Étonnamment, LIX s'est révélée adéquate pour l'ensemble de ces langues, sans qu'il soit nécessaire d'en modifier l'équation. Malheureusement, publiés en suédois, ces travaux vont rester longtemps inconnus des principaux chroniqueurs du domaine (Gilliland, 1972; Harrisson, 1980; Klare, 1974). Il faut attendre un communiqué d'Anderson (1981) pour que le monde anglo-saxon en prenne connaissance, avec un certain enthousiasme d'ailleurs.

Ce même Anderson (1983) va d'ailleurs proposer une version simplifiée de LIX, qu'il appelle RIX (Rate Index) et dont les prédictions sont exprimées en fonction des niveaux scolaires américains. Comme son nom l'indique, RIX prend la forme d'un ratio (Anderson, 1983, 496) :

$$RIX = \frac{\text{nombre de mots de plus de six lettres}}{\text{nombre de phrases}} \tag{2.34}$$

Ce ratio représente le taux de mots longs par phrase et prend indirectement en compte la longueur des phrases, ce qui en fait bien une formule à deux facteurs, un lexical et un syntaxique. Le recours à cette version simplifiée et plus pratique pour les éducateurs américains, est justifié par une corrélation de $0,99$ avec LIX.

Citons enfin Elley (1969), dont les résultats atypiques laissent songeur le chercheur en lisibilité. Tout d'abord, à l'instar de Björnsson (1983), elle se refuse à recourir au test de closure comme critère : elle ne l'estime pas plus efficace qu'un étalonnage manuel. Dès lors, elle évalue la complexité des textes qui forment ses cinq corpus de validation au moyen d'un groupe de juges. Ensuite, elle expérimente diverses variables sur son critère, en accord avec la démarche classique en lisibilité, mais en vient à rejeter cette approche au profit d'un facteur lexical unique : la fréquence moyenne des noms présents dans le passage à évaluer. En effet, cette variable seule est davantage corrélée avec son critère que chacun des autres indices envisagés, mais surtout, et c'est là le plus étonnant, elle se révèle également plus efficace que la combinaison d'un facteur lexical et syntaxique.

Son choix de ne prendre en compte que la fréquence des noms est dicté par les considérations suivantes. D'une part, elle estime que dénombrer le nombre d'absents d'une liste de référence constitue une approche assez grossière et qu'il est préférable de calculer la moyenne des fréquences, une procédure déjà employée par Lively et Pressey (1923), Lorge (1939) ou Forbes et Cottle (1953). D'autre part, poussant plus loin ses expériences, elle ob-

serve qu'il est à la fois plus efficace et plus économique (en termes d'efforts à fournir) de ne prendre en compte que les noms. Pour connaître la fréquence des noms, elle se reporte à la *New Zealand Council for Educational Research Alphabetical Spelling List*, publiée en 1961 et qui comprend 2700 mots classés en sept niveaux de fréquences, auxquels elle ajoute un huitième niveau constitué des 270 noms compilés par Holder en 1965. Sur la base de cette liste, elle peut ainsi associer la moyenne obtenue à un âge adéquat pour comprendre le texte analysé.

McLaughlin : Simple Measure of Gobbledygook

La même année que Bormuth (1969b) paraît un article de McLaughlin (1969) qui présente sa Simple Measure of Gobbledygook (SMOG), une formule qui pousse le désir de simplicité classique à l'extrême. Elle prend la forme suivante (McLaughlin, 1969, 643) :

$$RGL = 20 - \sqrt{X_2} \tag{2.35}$$

où :

RGL **:** le niveau scolaire pour lequel le texte est adapté ;

X_2 **:** le nombre de mots de plus de deux syllabes dans un échantillon de 30 phrases.

Dans cette équation, l'auteur semble ne recourir qu'à une seule variable explicative, inspirée de Gunning (1952). Or, en fixant le nombre de phrases de l'échantillon à trente, McLaughlin prend aussi en compte la longueur des phrases, puisque plus celles-ci sont longues et plus elles comportent de mots polysyllabiques [39]. Il modélise ainsi l'interaction entre la difficulté lexicale et la complexité syntaxique, ce qui avait déjà été fait par Wheeler et Smith (1954).

Il faut noter que McLaughlin (1969), s'il utilise aussi les leçons de McCall and Crabbs de 1961 comme critère, est plus exigeant en ce qui concerne le niveau de réussite requis. Il estime que seule la réussite totale au test (100% de bonnes réponses) indique une bonne compréhension. Il n'est dès lors pas étonnant que sa formule donne des résultats qui soient en moyenne plus élevés de deux niveaux scolaires que les autres formules.

Les formules des années septante

Après l'arrivée du test de closure en lisibilité avec Coleman et les études capitales de Bormuth qui entraînent toutes deux un renouveau de la discipline, le début des années septante est marqué par une grande variété de travaux aux orientations multiples. Afin d'organiser

39. Il s'agit de la même réflexion que celle opérée par Anderson (1983) pour RIX. Ce dernier reconnait d'ailleurs s'être inspiré des travaux de McLaughlin (1969).

leur présentation, nous les avons classés, autant que faire se peut, selon six axes qui soulignent leur caractéristique la plus saillante : l'accent mis sur la syntaxe, la conception de formules spécialisées, le *levelling* ou nivellement manuel, les approches par graphes, une certaine sensibilité aux théories cognitivistes ou, simplement, la volonté – héritée de la période classique – de produire une formule généraliste.

Focus sur la syntaxe : Un des efforts les plus intéressants en ce début de décennie est dû à Botel et Granowsky (1972). Ces deux chercheurs considèrent que la variable syntaxique communément utilisée, la longueur moyenne des phrases, ne constitue pas un facteur satisfaisant. Ils détaillent leurs critiques dans un article postérieur (Granowsky et Botel, 1974). Ils doutent par exemple que la complexité d'une phrase coordonnée telle que « Je suis allé au magasin et j'ai acheté du lait pour le diner. » soit deux fois supérieure à la difficulté des deux phrases suivantes : « Je suis allé au magasin. J'ai acheté du lait pour le diner. ». Inversement, si une phrase passive comporte le même nombre de mots que son équivalent actif, elle sera considérée de même difficulté alors que diverses études (Coleman, 1964; Gough, 1965) révèlent que la phrase passive est plus complexe à comprendre.

Ils proposent dès lors la « Syntactic Complexity formula » qui repose sur les théories de la grammaire générative et transformationnelle (Chomsky, 1957). Elle consiste à attribuer un poids, compris entre 0 et 3, aux divers types de structures syntaxiques et à en faire la moyenne pour obtenir la difficulté syntaxique d'un passage. Une technique qui rappelle fort les idées de Bormuth (1969a), mais n'est pas intégrée au sein d'une formule. Les auteurs insistent sur le fait que leur approche doit être considérée comme un procédé indicatif et non une mesure précise.

Dans la même veine, on trouve d'autres travaux qui s'attachent à mieux mesurer la difficulté syntaxique en s'inspirant également de divers travaux récents en linguistique. Brewer publie en 1972 une thèse intitulée *The effect of syntactic complexity on readability* où il expérimente diverses variables syntaxiques, mais doit en conclure que le meilleur prédicteur reste la longueur moyenne de la phrase. Dans une autre étude, MacGinitie et Tretiak (1971) parviennent à la même conclusion étonnante.

Enfin, Damerst (1972) décrit un nouvel indice de lisibilité qui semble assez convenu au premier abord, puisqu'il reprend les mêmes variables que Farr *et al.* (1951) : le nombre de mots polysyllabiques et le nombre moyen de mots par phrase. Toutefois, l'approche se base sur une hypothèse originale, à savoir que le nombre idéal de mots par phrase doit être compris entre 14 et 17. Damerst pense que les phrases trop longues gênent la lecture, mais aussi les phrases trop courtes, lesquelles produisent un style haché où le lecteur éprouve plus de mal à identifier les liens entre les idées. En résumé, il considère la relation entre la longueur des phrases et la difficulté comme parabolique, sans que cela ne complique

l'utilisation de son indice. Il suffit en effet de convertir les comptages obtenus pour les deux variables au moyen de deux tables avant d'additionner les nouvelles valeurs. Le résultat final est exprimé en fonction d'une échelle qui varie de 0 (le plus difficile) à 100 (le plus facile).

Formules spécialisées : Parmi les formules spécialisées, si l'un ou l'autre effort pédagogique subsiste, ce sont surtout les travaux à visée industrielle ou militaire qui dominent la décennie. Ainsi, Jablonski (1971, cité par Klare, 1974, 82-83) peaufine un profil spécialisé pour la compagnie Boeing, dont l'objectif est de situer la difficulté des manuels de maintenance sur une échelle de quatre niveaux. Il prend en compte les six facteurs suivants : la complexité des mots, la longueur des mots, la longueur des phrases, la quantité de connecteurs verbaux, de modifieurs et de références personnelles. Henshall (1971, cité par Klare, 1974, 83) s'intéresse quant à elle à la lisibilité de dictées sténographiques dans une thèse non publiée. Son étude présente un aspect atypique : elle est la première à concevoir une formule de régression dont chacune des deux variables sont elles-mêmes une formule de lisibilité, ici en l'occurrence l'Index Fog de Gunning et la formule de Farr-Jenkins-Paterson.

Au niveau militaire, Caylor *et al.* (1973) prennent part à un projet qui vise à évaluer les exigences en lecture des divers postes dans l'armée. Dans ce contexte, il développent FORCAST, une formule de lisibilité adaptée à des contenus militaires (caractérisés par l'emploi d'une terminologie particulière) et destinée à des adultes. Leur méthodologie est similaire à celle de Bormuth (1966) : ils rassemblent un ensemble de passages traitant de sujets aussi variés que possible [40], avant de leur appliquer la formule de Flesch (1948), ce qui leur permet de ne conserver qu'un corpus d'une douzaine d'extraits où les divers niveaux de difficulté sont représentés.

Ensuite, 395 recrues ont servi de sujets pour des tests de closure sur ces 12 passages et leurs résultats ont été transformés en *grades*. Les auteurs ont alors effectué une série de mesures de corrélations entre ce critère et quinze variables linguistiques pour n'en conserver qu'une seule. L'équation est la suivante (Caylor *et al.*, 1973, 15) :

$$X_1 = 20 - \frac{X_2}{10} \tag{2.36}$$

où :

X_1 **:** le niveau scolaire minimal auquel 50% des soldats lisant à ce niveau obtiendraient au moins 35% de bonnes réponses à un test de closure sur ce texte ;

X_2 **:** le nombre de mots monosyllabiques dans un échantillon de 150 mots.

Il faut noter que si le r calculé sur les données d'entraînement est élevé ($0,86$), FORCAST

40. Il s'agit de sujets tels que les armes légères, la mécanique, les techniques de soin ou encore la réparation de radars.

retombe à $0,77$ en validation croisée, faisant moins bien que les formules de Flesch (1948) ou de Dale et Chall (1948a) sur le même corpus. L'objectif de simplicité, au moins, est atteint.

Poursuivant cette voie, Kincaid *et al.* (1975) réutilisent la méthodologie de Caylor *et al.* (1973) afin d'adapter trois formules – ARI, le Fog Count de McElroy et la RE de Flesch – au contexte martial. Ils recourent à l'aide de 531 sujets pour étalonner 18 extraits de manuels militaires selon l'échelle des *grades* américains. Les trois formules obtenues sont :

$$\begin{aligned} \text{GL (ARI)} &= -27,4 + 0,4\,(\text{mots par phrase}) + 6\,(\text{traits par mots}) \\ \text{GL (Fog Count)} &= \frac{((X_2 + 3X_3) \diagup X_4) - 3}{2} \\ \text{GL (Flesch)} &= -15 + 0,4\,\text{ASL} + 12\,\text{ASW} \end{aligned} \tag{2.37}$$

où :

GL **:** le niveau scolaire recalculé prédit pour le texte donné ;

X_2 **:** le nombre de mots d'une ou deux syllabes dans un échantillon de 100 mots ;

X_3 **:** le nombre de mots de plus de deux syllabes dans le même échantillon ;

X_4 **:** le nombre de phrases dans l'échantillon ;

ASL **:** le nombre moyen de mots par phrases ;

ASL **:** le nombre moyen de syllabes par mots ;

C'est surtout la dernière de ces formules qui a connu le succès, sous la dénomination Flesch-Kincaid, puisque c'est elle qui sera employée pour contrôler la difficulté de lecture de certains manuels de l'armée.

En parallèle aux travaux financés par l'armée, Harris et Jacobson (1973, 1974) publient cinq formules [41] pour un tout autre domaine : les matériaux à destination des primaires. Certaines d'entre elles sont spécifiques à la première moitié des primaires, tandis que les autres sont plutôt adaptées à la seconde moitié du cycle. Elles ont été développées en profitant des capacités accrues de traitement offertes par les ordinateurs. Les auteurs sont ainsi capables d'analyser 661 passages d'environ 200 mots extraits de *readers*, ces livres simplifiés à destination des enfants. Le niveau scolaire de chacun de ces *readers* étant spécifié, c'est ce classement qui sert de critère pour étalonner les passages.

Les travaux de Harris et Jacobson présentent plusieurs aspects intéressants envisagés en

41. À nouveau, nous renvoyons le lecteur intéressé par le détail des formules et de leur niveau de validité et de fiabilité à l'article de Harris et Jacobson (1974, 3).

détail dans le Chapitre 3. Parmi ceux-ci, le plus novateur est certainement la création d'une nouvelle variable référençant un ensemble de motifs orthographiques. Elle repose sur les expériences de Jacobson (1974), lequel montre que le degré de correspondance entre les graphèmes et les phonèmes, mesuré au moyen de 37 motifs orthographiques particuliers, est hautement corrélé ($R = 0,92$) avec les difficultés à la lecture.

Harris et Jacobson (1976) poursuivent ensuite leurs recherches selon une autre approche. S'opposant aux conclusions de Bormuth (1966) pour qui une seule formule suffit pour l'ensemble des niveaux scolaires, ils conçoivent des formules particulières à la sixième primaire, la troisième secondaire et la rhétorique. Pour chacune d'elles, ils utilisent comme critère les résultats de plusieurs milliers de sujets à des tests de compréhension portant sur quelques dizaines de passages issus des leçons de McCall et Crabbs. Harris et Jacobson optent également pour des formules informatisées, qui comprennent plusieurs variables et obtiennent des très hauts taux de corrélation (6e primaire : $0,89$; 3e secondaire :$0,88$ et 6e secondaire : $0,96$). Toutefois, l'article en question ne rapporte pas les équations. On sait juste que l'équation spécifique à la rhétorique comprend douze variables : le pourcentage de mots absent de la liste de Harris et Jacobson (1972) ; le nombre moyen de mots par phrase et dix motifs orthographiques.

Enfin, Hull (1979, cité par Klare (1984, 696)) inscrit sa réflexion sur la difficulté des textes techniques dans la lignée de FORCAST et des formules de Kincaid *et al.*. Il cherche cependant à remplacer la longueur des mots par une autre variable et envisage l'effet de certaines classes de mots. Il part du principe qu'une augmentation du nombre d'adjectifs et d'adverbes situés devant un nom diminue la lisibilité d'un matériel. Pour valider cette hypothèse, il reprend les cinq passages du critère de Kincaid *et al.* dont il propose également deux versions modifiées. Il obtient effectivement une corrélation négative ($-0,664$) entre ce critère et sa nouvelle variable et en dérive une formule à deux variables (la seconde est la longueur de la phrase) qui explique 48% de la variance ($R = 0,69$). Il en conclut que sa nouvelle variable, bien que légèrement moins prédictive que le nombre moyen de syllabes par mots, lui est toutefois préférable, car elle entretient un rapport causal avec la difficulté et peut dès lors être manipulée pour rendre un texte plus accessible.

Le levelling : Tandis que les années 70 s'écoulent, les détracteurs de la lisibilité expriment des critiques de plus en plus vives (que nous détaillons à la section 2.2.6), au point que certains chercheurs en reviennent à une autre forme d'évaluation de la difficulté : le *levelling*. Discuté par Klare (1963) et Fry (2002), il s'agit d'une approche déjà envisagée par Thorndike (1910, cité par Singer, 1975, 256) ou Ojemann (1934) et qui consiste à restituer à l'humain la tâche d'évaluation, mais en guidant son jugement à l'aide d'un cadre de référence.

Ainsi, Morris (1972) propose le Clear River Test, qui ressemble à une généralisation de l'idée de brouillard (*fog*) défendue par Gunning (1952). En effet, il identifie quatre variables lexicales et syntaxiques classiques, pour lesquelles il définit une valeur qu'il vaut mieux ne pas dépasser lorsqu'on rédige. Chacune des variables est envisagée séparément. Morris ne les combine pas en un seul indicateur. De plus, il ne rapporte pas de données de développement, ni de validation. On se situe plutôt du côté de la simplification de textes que de la lisibilité.

Plus connus sont les travaux de Carver et de Singer. Tous deux partagent la conviction que l'humain peut être aussi précis que des formules, non seulement pour ordonner un ensemble de passages en fonction de leur difficulté, mais aussi pour situer des textes sur une échelle absolue telle que, par exemple, les années scolaires. Dans un premier temps, Carver (1974a) définit l'échelle Reading-Input Difficulty Estimate (RIDE), qui consiste simplement à compter le nombre de lettres par mots. Il effectue plusieurs études de validation, sur les passages de Miller et Coleman (1967) et de Bormuth (1969a), et obtient, pour RIDE, des coefficients de corrélation similaires à ceux d'autres formules de lisibilité. Toutefois, il n'est pas convaincu par la capacité de ce type de formules à traiter certains phénomènes tels que la complexité conceptuelle. Carver (1974a, 1975) se tourne alors vers le *levelling* et définit la *Rauding Scale*.

Cette échelle comporte six textes considérés comme représentatifs des *grades* 2, 5, 8, 11, 14 et 17. Carver considère que des individus peuvent fiablement évaluer la lisibilité d'un texte en le comparant à ces étalons, pour peu qu'ils aient préalablement fait la preuve de leur habilité à expertiser des textes en situant correctement cinq passages de référence. La valeur de la *Rauding Scale* pour un texte donné s'obtient en faisant la moyenne de trois jugements humains et en discrétisant la valeur obtenue à l'aide d'une table de conversion. Carver (1974a, 1975) montre, à l'aide de deux expériences, que sa méthode est légèrement plus fiable que des formules classiques telles que celles de Flesch (1948) ou Dale and Chall (1948).

Un autre représentant notable de cette tendance au *levelling* est Singer (1975). Il publie une approche intitulée Singer Eyeball Estimate of Readability (SEER) qui ressemble à celle de Carver. Elle consiste aussi à rapprocher un passage dont la difficulté est inconnue à un autre extrait faisant partie d'un corpus de référence et dont le niveau scolaire est connu. L'amplitude de son échelle est toutefois plus réduite – elle s'étend du milieu de la première primaire à la fin de la première secondaire – et, par conséquent, plus précise. En effet, elle comprend pas moins de deux extraits par *grade*. Singer en vérifie la validité au moyen de 32 juges et conclut qu'elle est aussi précise que le graphe de Fry (1968).

On peut se demander qui, de Carver ou de Singer, obtient les meilleurs résultats sur un même

corpus de test. Une première comparaison de Froese (1979, cité par Duffelmeyer, 1982, 5), réalisée à l'aide de 42 juges, conclut en faveur de Singer. Toutefois, l'étude souffre de deux faiblesses méthodologiques : d'une part, ce sont les passages choisis comme référence par Singer qui composent le corpus de test, d'autre part, les juges n'ont pas passé le test de qualification défini par Carver. Dès lors, Duffelmeyer (1982) entreprend une nouvelle comparaison exempte de ces deux défauts. À l'inverse, il trouve que ce sont les passages évalués à l'aide de la *Rauding Scale* qui obtiennent les résultats les plus proches des estimations fournies sur ces mêmes passages par les formules de Dale et Chall (1948a) et de Spache (1960).

Les approches par graphes : Les approches par graphes, dont la vocation première reste la poursuite de la simplicité, connaissent trois développements durant les années septante. Mugford (1970), un anglais, établit un graphe qui évalue la difficulté directement en fonction de l'âge des lecteurs selon un spectre qui s'étend de cinq ans et demi à quinze ans. Il reprend les deux mêmes variables que Fry (1963), auxquelles il ajoute une mesure de la redondance.

Fry (1977) lui-même, dont le graphe publié précédemment est devenu très apprécié et utilisé, étend son amplitude en ajoutant des repères pour le niveau du collège. Il effectue cette extension par simple extrapolation, une démarche qui manque d'une certaine rigueur méthodologique, comme il le souligne lui-même.

La même année, Raygor (1977, cité par Baldwin et Kaufman, 1979) sort un graphe proche de celui de Fry, où le décompte des syllabes est remplacé par le nombre de mots de six lettres ou plus, une variable déjà envisagée par Gunning (1952). Cette modification est motivée par le désir de gagner du temps lors de l'analyse et de rendre celle-ci plus simple. Deux arguments qui sont légitimés par les expériences de Baldwin et Kaufman (1979). Elles révèlent que l'utilisation du graphe de Raygor entraîne bien un gain de temps d'environ 25% et que, d'autre part, le décompte des syllabes engendre bien un nombre supérieur d'erreurs.

Formules généralistes : Les années septante donnent naissance à de rares formules généralistes, parmi lesquelles on peut citer celle de Coleman et Liau (1975). Ces auteurs partent du constat suivant : pour appliquer une formule entièrement automatisée à un texte, il est nécessaire de transcrire celui-ci sur une carte perforée, une opération qui requiert autant d'efforts que l'utilisation d'une formule manuelle. Ils suggèrent d'employer à la place un scanner optique. À l'aide des 36 passages calibrés par Miller et Coleman (1967), ils conçoivent la formule suivante (Coleman et Liau, 1975, 284) :

$$X_1 = 141,8401 - 0,21459\,L + 1,079812\,S \tag{2.38}$$

dont les variables sont :

X_1 **:** la proportion de réponses correctes qu'obtiendrait un étudiant de l'enseignement supérieur à un test de closure portant sur le texte ;

L **:** le nombre de lettres par 100 mots ;

L **:** le nombre de phrases dans 100 mots.

Ils rapportent un R atteignant $0,92$ et fournissent également une équation permettant d'exprimer les résultats de leur formule en termes de niveaux scolaires. Cette équivalence est obtenue à l'aide des 26 passages de leur corpus qui sont tirés des leçons de McCall et Crabbs. Elle est donc à considérer avec une certaine méfiance étant donné ce qui a été dit à leur sujet.

L'année suivante, Liau *et al.* (1976) publient une révision des premiers travaux de Coleman (datant de 1965) dont ils cherchent à améliorer deux dimensions : la qualité du critère et la sélection des indices de la difficulté. Ils dérivent ainsi plusieurs centaines de formules à partir de 32 variables explicatives et du corpus de Miller et Coleman (1967). Parmi celles-ci, ils en retiennent quatre [42] dont le R varie de $0,9$ à $0,94$.

Les premiers efforts cognitivistes : Après les expérimentations de Bormuth, certains chercheurs en viennent à penser que tous les facteurs envisageables ont été étudiés. D'autres sont convaincus du contraire et explorent quelques variables d'inspiration cognitiviste. Cette tendance, assez atypique, laisse déjà entrevoir les chamboulements qu'apportera le vue cognitiviste en lisibilité au cours de la période suivante (*cf.* Section 2.2.6).

Ainsi, Cohen (1975, cité par Klare, 1984, 695) s'essaie à compter les mots abstraits, comme l'avait fait auparavant Flesch (1950). Il les définit toutefois à l'aide de règles différentes : il ne se base plus sur les catégories de discours, mais sur leur présence dans une liste de mots abstraits ou, à défaut, sur la présence de certains suffixes ou finales de mots. Six ans plus tard, Colby (1981, cité par Klare, 1984, 695) poursuit cette idée, sous la forme d'un ratio des mots extensionnels et intentionnels. Toutefois, aucun des deux auteurs ne propose une formule en soi.

Parallèlement à ces deux mesures de l'abstraction, Williams, Siegel, Burkett et Groof présentent en 1977 *The Air Force transformational formula* que DuBay (2004, 32) décrit comme l'étude qui est peut-être la plus ambitieuse parmi celles qui ont tenté d'adapter les théories cognitivistes sous une forme quantitative afin de réaliser un outil de mesure de la difficulté. Ils incluent dans leur formule huit variables, dont quatre sont de nature syntaxique (complexité transformationnelle, poids structural...) et quatre touchent à la structure de l'intellect. Ces quatres dernières sont absolument inédites à l'époque, mais leur intégration ne permet que d'atteindre un R de $0,601$ sur un corpus de 14 passages, dont le niveau de complexité a été évalué à l'aide d'un test de closure. Ce premier essai n'apparaît donc guère convaincant.

42. Pour le détail des équations, se reporter à la page 384 de Liau *et al.* (1976). On peut noter que la seconde de ces formules correspond à celle publiée par Coleman et Liau (1975).

Les formules de lisibilité dans le monde francophone

À la même époque, le monde francophone est lui aussi stimulé par l'arrivée du test de closure en lisibilité. De Landsheere s'en fait l'un des principaux promoteurs à travers diverses publications (de Landsheere, 1972, 1978, 1982). Il montre que le test de closure est un moyen efficace de discriminer les différents niveaux scolaires de l'enseignement belge et défend dès lors son usage comme mesure de la lisibilité des textes.

Mais c'est l'un de ses disciples, Georges Henry qui propose les premières formules spécifiques au français L1 dans sa thèse de doctorat publiée en 1973. Pour ce faire, il constitue un corpus de 60 textes tirés de divers manuels scolaires de la première primaire à la fin de l'enseignement secondaire. La difficulté de chaque passage est évaluée au moyen de tests de closure, qui ont été administrés à trois groupes de 60 sujets. Ces groupes sont respectivement constitués d'élèves de la 5e et 6e primaire, de la 2e et 3e secondaire et de la 5e et 6e secondaire et chacun d'eux donne naissance à une famille de formules qui lui est propre. En ce sens, Henry diverge de Bormuth (1966), qui estimait qu'une seule formule était suffisante pour l'ensemble des niveaux scolaires. Par contre, comme lui, il propose des formules différentes selon leur contexte d'utilisation. Henry (1975, 94-95) définit ainsi trois formules manuelles, trois automatisées et trois dont la validité est maximale et qui combinent les huit meilleures variables parmi les 116 testées. Nous ne rapportons ici que les formules automatisées, dans leur version revue de 1977 [43] (Henry, 1979, 57) :

$$\begin{aligned}
Logit_{10}\, P_{5-6} &= -(0,127\,\text{TTR}) - (0,085\, logit_{10}\,\text{PPD}) - (0,004\,\text{MP}) \\
&\quad -(0,519\, logit_{10}\,\text{AG}) + (0,136\, logit_{10}\,\text{DEXGU}) \\
&\quad +(0,188\, logit_{10}\,\text{NOMCONC}) - 0,219 \\
Logit_{10}\, P_{8-9} &= -(0,863\,\text{TTR}) - (0,024\, logit_{10}\,\text{PPD}) - (0,003\,\text{MP}) \\
&\quad -(0,305\, logit_{10}\,\text{AG}) + (0,069\, logit_{10}\,\text{DEXGU}) \\
&\quad +(0,133\, logit_{10}\,\text{NOMCONC}) + 0,658 \\
Logit_{10}\, P_{11-12} &= -(0,973\,\text{TTR}) - (0,021\, logit_{10}\,\text{PPD}) - (0,003\,\text{MP}) \\
&\quad -(0,357\, logit_{10}\,\text{AG}) + (0,026\, logit_{10}\,\text{DEXGU}) \\
&\quad +(0,083\, logit_{10}\,\text{NOMCONC}) + 0,669
\end{aligned} \tag{2.39}$$

où :

$Logit_{10}P_{5-6}$ **:** le logit [44] (en base 10) du pourcentage moyen de réussite des étudiants de 5e et 6e primaire au test de closure sur un texte donné ;

43. La principale différence entre la formule automatisée de 1973 et celle de 1977 est l'ajout du pourcentage de noms concrets dont Henry a établi une liste.

44. Le logit est défini de la façon suivante :

TTR **:** la proportion de types par rapport au nombre de tokens (Type Token Ratio) ;

PPD **:** le pourcentage de pronoms personnels de dialogue ;

MP **:** le nombre moyen de mots par phrases ;

AG **:** le pourcentage d'absents de la liste de Gougenheim *et al.* (1964) ;

$DEXGU$ **:** le pourcentage d'indicateurs de dialogue ;

$NOMCONC$ **:** la proportion de mots concrets ;

La validité de ces trois formules est relativement élevée : $R = 0,89$ pour P_{5-6} ; $R = 0,89$ pour P_{8-9} et $R = 0,81$ pour P_{11-12}.

L'autre acteur majeur du domaine, François Richaudeau, s'intéresse d'abord aux processus de lecture et à la mémoire (Richaudeau, 1969). D'une part, il analyse les mouvements de l'oeil et en conclut que la typographie n'a guère d'influence sur la lisibilité d'un texte. D'autre part, il étudie l'effet de la mémoire dans le processus de lecture et estime qu'un texte lisible est assimilable à un texte plus ou moins bien retenu. On peut donc utiliser comme critère de compréhension le nombre de mots retenus après lecture.

Dans des travaux postérieurs (Conquet et Richaudeau, 1973), il présente une formule de lisibilité reposant sur cette hypothèse. Il y définit la *sous-phrase* comme l'unité linguistique de la lisibilité, car celle-ci est plus aisément définissable et constitue une suite de mots autonomes. Il note que les phrases courtes ou prévisibles et les mots usuels favorisent la mémorisation et doivent dès lors être privilégiés. Il propose donc la formule suivante (Conquet et Richaudeau, 1973, 15) :

$$X_1 = A + B - C \tag{2.40}$$

où :

X_1 **:** Le nombre de mots retenus d'une sous-phrase ;

A **:** une valeur obtenue sur un tableau à double entrée qui combine la longueur moyenne des mots et des phrases ;

B **:** vaut 2 quand le début de la phrase est une formule courante d'affirmation, 0 dans le cas contraire ;

C **:** vaut 2 quand le nombre de mots qui sépare le sujet et le verbe est supérieur à 10 mots, 0 dans le cas contraire ;

$$logit_{10}\, x = \log_{10} \frac{x}{1-x}$$

.

Richaudeau poursuit ensuite ses réflexions, nourries de ses expériences en psychologie de la lecture. Il critique les formules de lisibilité classiques (Richaudeau, 1976), trop dépendantes à ses yeux du test de closure qui ne mesurerait que la redondance d'un texte, c'est-à-dire son conformisme. En conséquence, Richaudeau (1978) en vient à développer un nouveau concept : l'**efficacité** linguistique, qu'il distingue de la lisibilité. Un texte peut être lisible parce qu'il est redondant (son *entropie* est faible), ce qui n'est pas le cas d'un texte efficace, qui est toujours porteur d'informations nouvelles et laisse une empreinte dans l'esprit du lecteur.

Sur la base de ces études, Richaudeau (1979, 1981, 1984) défend sa formule d'efficacité, qui privilégie les facteurs syntaxiques de la phrase et tient compte du fonctionnement de l'esprit humain. Elle utilise toujours comme critère la quantité de mots d'une phrase stockés par la mémoire à court terme. Dès lors, le taux d'efficacité linguistique d'un texte correspond au « rapport entre le nombre de mots exacts dont se souvient le lecteur et le nombre total de mots de cette phrase » (Richaudeau, 1979, 139). La formule a été conçue à l'aide des résultats de 47 sujets sur 37 phrases qui ont été corrélés avec diverses variables. Richaudeau obtient, sans recourir à la régression linéaire multiple, une équation dont les coefficients ont été définis par essais et erreurs (Richaudeau, 1979, 146) :

$$\begin{aligned} R &= \frac{S * P}{M} \\ \text{où} \quad S &= 15,3 + b + 0,5c - 0,5d - 4e - 6f - 6g \end{aligned} \tag{2.41}$$

où :

- R **:** l'indice d'efficacité linguistique ;
- P **:** un coefficient de pondération, qui dépend de la qualité du lecteur et du type de phrase (affective, courante ou abstraite) ;
- M **:** nombre de mots totaux dans la phrase ;
- S **:** le nombre standard de mots mémorisés ;
- b **:** le nombre de mots outils indicateurs (car, comme, dont, en, où, etc.) ;
- c **:** le nombre de répétitions ;
- d **:** le nombre de termes d'énumération (non compris le premier) ;
- e **:** vaut 1 si le verbe est en dernier mot, 0 sinon ;
- f **:** vaut 1 s'il y a un enchassement de plus de 15 mots, 0 sinon ;
- g **:** vaut 1 quand la structure est très monotone (pas de critères précis décrits par Richaudeau).

Cette formule, dont le coefficient de corrélation n'est pas rapporté, ne connaît pas le même succès que celle d'Henry, sans doute parce qu'elle est moins pratique. En effet, tandis

qu'Henry se pose en dépositaire de la tradition anglo-saxonne fortement empirique, Richaudeau correspond davantage à ce que peut donner le mariage de la sensibilité française et de la lisibilité. Il se refuse à reprendre telle quelle une méthodologie vieille de près de soixante ans et s'efforce de faire le lien entre la lisibilité et les nouvelles connaissances des processus cognitifs de l'homme, au détriment, toutefois, d'une certaine applicabilité de ses travaux.

Conclusion

Cette quatrième période dans l'histoire de la lisibilité a vu paraître nombre de recherches originales et dynamiques qui ont ébranlé les paradigmes classiques du domaine, tout en ouvrant de nouveaux horizons. Coleman a montré que l'on gagne à abandonner le test de compréhension au profit du test de closure, ce qui permet d'obtenir de meilleurs taux de corrélation avec les prédicteurs de la difficulté [45]. Bormuth (1966) a théorisé les intuitions de certains de ses prédécesseurs et a démontré que la relation entre la difficulté et les variables linguistiques n'est pas nécessairement linéaire. Il a également défendu la conception de formules plus complexes et validées sur un corpus différent de celui qui a servi à l'entraînement.

Par ailleurs, l'automatisation croissante du domaine a entraîné des modifications sensibles dans les approches. Le temps de comptage des variables importe de moins en moins ; c'est désormais la possibilité de spécifier les normes de comptage à une machine qui devient l'élément prépondérant. Cela conduit plusieurs chercheurs à se limiter à des facteurs de surface, plus aisément explicitables, mais dont le dénombrement via des routines informatisées conduit souvent à simplifier les définitions des unités linguistiques employées. Une phrase est délimitée par deux points, un mot par deux blancs. On ignore certains phénomènes linguistiques tels que les mots composés, les sous-phrases, les abréviations, etc.

Dès lors, avec l'arrivée des travaux de psychologie cognitive en lisibilité, déjà annoncée par les études de Cohen, Colby, Williams *et al.* ou de Richaudeau (1979) pour le français, des critiques de plus en plus vives vont être adressées à la lisibilité, comme nous allons le découvrir dans la section suivante.

45. Plusieurs auteurs (Elley, 1969; Klare, 1984) estiment que ces résultats étonnamment élevés s'expliquent par la nature du critère. Si celui-ci comprend un grand nombre de passages aux contenus divers et à la difficulté variée, le R obtenu sera élevé. Pour peu qu'à ces circonstances favorables s'ajoute le recours à des sujets dont les compétences en lecture divergent fortement, il est alors possible de maximiser le R et d'obtenir des résultats dépassant le $0,9$. C'est pourquoi Klare (1984, 706) recommande de privilégier l'erreur standard d'estimation au coefficient de corrélation multiple lors de l'évaluation d'un modèle.

2.2.6 Critiques et élans structuro-cognitivistes (1979 à 1989)

Introduction

À partir des années 70, des critiques de plus en plus vives s'élèvent à l'encontre des formules de lisibilité, lesquelles oublieraient de considérer certaines caractéristiques textuelles essentielles ou abstrairaient les propriétés du texte de la situation de lecture et du lecteur. Ces griefs trouvent leur origine dans le renouveau des travaux en psychologie cognitive de la lecture, mais aussi – et il ne faut pas l'oublier – dans l'autocritique à laquelle se livrent nombre de chercheurs en lisibilité (Harris et Jacobson, 1979; Jacobson *et al.*, 1978b; Klare, 1984). La confiance dans la méthodologie classique s'étant effritée, cela se traduit par une critique systématique de cette approche.

Simultanément, cette critique assez radicale s'accompagne d'une exploration de voies parallèles, parmi lesquelles domine l'approche structuro-cognitive. Elle se nourrit de la révolution cognitiviste que Spivey (1987, 169) décrit ainsi :

> *It is difficult to determine when constructivism began to take hold in the United States. Sometime in the early 1970s cognitive psychologists moved beyond words and sentences to texts as they studied the mental processes involved in human memory and understanding. Some psychologists were beginning to argue constructivist notions : that people, rather than texts, carry meaning and that linguistic inputs merely provide "cues" that readers use along with their knowledge of the world to construct meaning.*

Cette nouvelle vision de la lecture modifie sensiblement les orientations des chercheurs en lisibilité, en particulier au niveau des facteurs utilisés pour prédire la difficulté. Pour résumer, on peut dire que les indices lexico-syntaxiques sont délaissés au profit de variables structuro-cognitives, telles que la cohésion, l'organisation textuelle, la quantité d'inférences, etc.

Dans cette section, nous commençons par détailler les critiques faites par les cognitivistes à la lisibilité classique, avant de proposer un parcours des travaux en lisibilité qui s'appuient sur ce paradigme structuro-cognitive. Nous terminons avec les études qui ont échappé à cette mode et se situent dans la lignée des travaux précédents.

Les critiques des structuro-cognitivistes

Comme nous l'avons déjà dit, la fin des années 70, et surtout les années 80, voient augmenter le nombre de critiques envers la lisibilité, dont certaines sont particulièrement virulentes. Citons, parmi les plus acerbes, les articles de Selzer (1981), *Readability is a four-letter word*,

et de Duffy (1985), *Readability formulas : What's the Use ?*, dont les titres seuls suffisent à se faire une idée du contenu. Même si les articles de ce genre abondent à l'époque, ils recourent souvent aux mêmes arguments, qui en deviennent parfois caricaturaux.

Prenons comme exemple de cette mauvaise foi – qui prend parfois des allures d'oubli sélectif des travaux antérieurs – quelques déclarations tirées de Bruce *et al.* (1981, 3-4). Les auteurs avancent ainsi : « Many of the hundreds of formulas in existence were validated only in terms of earlier formulas ». Nous avons vu au cours de ce chapitre combien cette déclaration est spécieuse : si quelques études mineures se sont effectivement limitées à ce type de validation, la plupart des chercheurs en lisibilité sont soucieux de la qualité de leur critère. Bruce *et al.* (1981, 4) poursuivent : « The early formulas, in turn, were validated using the McCall-Crabbs Standard Test Lessons in Reading ». Cette réflexion est évidemment suivie d'une référence à l'étude de Stevens (1980), afin de porter le coup de grâce. En trois phrases, toute validité est déniée à l'entièreté des travaux en lisibilité. C'est, il nous semble, faire fi un peu vite de la rigueur de chercheurs tels qu'Ojemann (1934), Gray et Leary (1935), Miller et Coleman (1967), Bormuth (1966) ou Henry (1975), entre autres.

Il est toutefois évident que certaines des critiques exprimées à l'époque sont pertinentes, même si elles méritent d'être débattues. Nous les avons résumés au sein des cinq points suivants :

Le manque d'ancrage théorique : Plusieurs chercheurs (Bruce *et al.*, 1981; Kemper, 1983; Kintsch et Vipond, 1979; Redish et Selzer, 1985) reprochent aux formules de lisibilité leur manque d'ancrage théorique, qu'il faut plutôt interpréter comme une trop faible prise en compte des travaux en psychologie cognitive. Ainsi, Bruce *et al.* (1981, 1) estiment que les formules « ignorent ou violent la plupart des connaissances actuelles à propos de la lecture et du processus de lecture ». Pour leur part, Kintsch et Vipond (1979, 335) y voient là l'explication d'une tendance qu'ils déplorent : « en l'absence d'une théorie de la structure textuelle et de son traitement, les chercheurs sélectionnent dans le texte une série de facteurs de surface qui sont assez évidents ».

Cette première critique est toutefois largement rejetée par Chall et Dale (1995, 81-82) qui insistent sur le fait que les chercheurs en lisibilité se sont nourris d'apports théoriques divers issus de la psycholinguistique, des théories du langage, des études sur le développement des capacités de lecture ou de l'intelligence, ou encore des travaux sur la compréhension à la lecture. Notons que Bormuth (1966, 94) va dans le même sens lorsqu'il estime que les progrès effectués en lisibilité au cours des années 60 sont notamment liés au développement d'une nouvelle discipline, la psycholinguistique, qui organise diverses découvertes en un ensemble de théories utilisables.

D'ailleurs, lorsqu'on analyse l'argumentaire qui fonde cette critique, on se rend compte qu'il cache en fait deux autres remarques. Si les formules de lisibilité manquent d'assises théoriques, c'est plutôt parce qu'elles ne prennent ni en compte les nouvelles dimensions textuelles récemment découvertes, ni le fait que la lecture soit le résultat d'un processus interactif. On touche là aux deux critiques majeures des cognitivistes.

Les formules sont incapables de prendre en compte des variables plus complexes : S'il y a bien une objection qui fait l'unanimité, c'est l'incapacité de la quasi-totalité des formules à considérer autre chose que les facteurs de surface. Cette limitation, déjà reconnue par les chercheurs en lisibilité (Dale et Chall, 1949; Klare, 1963), est reprise par de nombreux cognitivistes (Bruce *et al.*, 1981; Kemper, 1983; Kintsch et Vipond, 1979; Redish et Selzer, 1985; Selzer, 1981), qui détaillent les dimensions textuelles oubliées. Pour Zakaluk et Samuels (1988, 122), les formules de lisibilité négligent des facteurs aussi importants que la cohésion des idées ou l'organisation macrotextuelle. En effet, il est possible de mélanger les phrases d'un texte et d'obtenir le même score de difficulté [46]. Pour Marshall (1979), les formules classiques ne mesurent pas véritablement le sens des textes, puisqu'intervertir les mots d'une phrase n'engendre aucune différence de lisibilité. Pour Kemper (1983), c'est la charge inférentielle du texte qui devrait être prise en considération, etc...

Plusieurs auteurs se sont toutefois inscrits en faux contre cette condamnation des variables classiques. Ainsi, Chall (1988, 10) juge que les cognitivistes des années 80 oublient un peu vite que ces mêmes limitations étaient déjà exprimées par divers chercheurs des « Beginning years », dont on tend à sous-estimer la finesse des découvertes. Pour sa part, Boyer (1992, 6) rappelle qu'il « est permis de croire qu'il existe des affinités entre les caractéristiques superficielles du texte mesurées par les formules et ces autres caractéristiques » (à savoir ici, la cohésion et la cohérence). Enfin, comme nous en avons discuté dans le Chapitre 1, la part réelle qu'occupent les éléments d'organisation textuelle sur la compréhension écrite reste discutée et a sans doute été surestimée au cours des années 80 et 90.

Les formules classiques sont basées uniquement sur des facteurs textuels : Le corolaire de la remarque précédente est bien résumé par Bruce *et al.* (1981, 2) : « Because the formulas are measurements based on a text isolated from the context of its use, they cannot reflect such reader-specific factors as motivation, interest, competitiveness, values, and purposes ». Autrement dit, ignorer le caractère interactif du processus de lecture revient à considérer que la population de lecteurs auxquels s'adresse un texte est homogène, et que la situation de lecture est unique. Divers cognitivistes (Kemper, 1983; Kintsch et Miller, 1984; Kintsch et Vipond, 1979; Marshall, 1979; Redish et Selzer, 1985) ne peuvent accepter ces hypothèses. Ils estiment que le niveau de motivation du lecteur, ses objectifs, ses connaissances dans le

46. Notons que Chall (1958) faisait déjà cette remarque bien auparavant.

domaine couvert ou le temps qu'il investit modifient considérablement la compréhension d'un même texte. Par conséquent, les formules sont inutilisables en situation réelle.

C'est oublier un peu vite que les formules classiques prennent bien en compte l'interaction texte-lecteur au travers des tests utilisés pour étalonner le corpus d'entraînement (Chall et Dale, 1995, 97), ce que reconnaîtront par la suite Kintsch et Miller (1984). C'est également pécher par excès de relativisme, comme y insiste Mesnager (2002, 31) : « Cette manière de mettre en avant d'emblée l'interpénétration des facteurs, qui confine à une revendication de "relativité générale", pourrait aboutir à une implicite et délétère idée qu'aucun texte n'est plus difficile qu'un autre ». En réalité, l'adresse du chercheur en lisibilité consiste à se positionner adéquatement sur le continuum qui va de l'universalité à la prise en compte des particularismes. Si l'approche cognitiviste se situe clairement du côté des particularismes, la lisibilité – dont les objectifs sont plus pragmatiques – aspire à l'universalité, mais ne doit pas oublier qu'elle ne peut tenir un discours prédictif que sur des lecteurs et un contexte suffisamment similaires à ceux qui ont été utilisés pour sa conception.

Problème de généralisation : Avec ces réflexions, nous anticipons sur la quatrième critique émise à l'encontre de la lisibilité : sa tendance à généraliser ses résultats au-delà de la population modélisée. L'un des exemples les plus patents de cette déviance est dénoncé par Selzer (1981) et Redish et Selzer (1985) : nombre de formules, dont celles de Flesch (1948) ou de Gunning (1952), ont été développées sur un corpus de textes pour enfants alors qu'elles se prétendent adaptées à des adultes. Il y a évidemment là un problème de généralisabilité que nous avons également souligné précédemment dans ce chapitre.

Redish et Selzer (1985, 3) signalent un autre abus : la tendance de certains utilisateurs des formules, dont entre autres l'État américain, à les appliquer à des situations ou des types de textes pour lesquels elles n'ont pas été prévues. Cette pratique est notamment dommageable parce que la plupart des formules ont été développées sur la base du critère qui considère qu'un taux de 50% de bonnes réponses à un test de compréhension suffit. Si ce seuil est sans doute pertinent dans un contexte éducatif, où l'objectif d'un passage est de développer les compétences des élèves, il apparaît nettement moins valide lorsqu'il est appliqué à un contexte professionnel où la compréhension doit être la plus complète possible.

Ce problème de généralisation abusive est plus critique qu'il n'y paraît. En effet, on peut penser qu'il suffit d'utiliser les formules de lisibilité en veillant à éviter les travers que nous venons de décrire. Malheureusement, les conditions qu'une situation doit remplir pour qu'on puisse lui appliquer une formule donnée sont nombreuses. Bruce *et al.* (1981, 9-10) les listent avant de conclure qu'ils ne voient pas de situations réelles où l'ensemble de ces conditions sont respectées.

Impossibilité d'utiliser les formules pour la rédaction : Enfin, l'attaque la plus vive envers les formules de lisibilité concerne leur usage lors de la rédaction, ce que l'on désigne sous le terme de *writability* ou simplification de textes. Bruce *et al.* (1981) expliquent que les auteurs ont tendance à écrire en fonction des formules, c'est-à-dire en utilisant des phrases courtes et des mots simples. Selzer (1981) explique que les rédacteurs oublient souvent que les variables prises en compte dans les formules ne constituent pas des causes de la difficulté, mais uniquement des indices qui lui sont corrélés. Par conséquent, il arrive que des phrases réécrites en fonction de ces variables, et estimées plus simples par une formule classique, soient en fait plus complexes, comme le montre une expérience de Davison et Kantor (1982).

Selzer (1981) note également que les formules de lisibilité ont fortement influencé l'enseignement stylistique au sein duquel les problématiques interphrastiques ont été négligées, de même que certains aspects phrastiques tels que le rythme, les stratégies de coordination et de subordination, la recherche de la variété, etc. On a eu tendance à penser qu'écrire avec des phrases courtes et des mots simples suffisait à produire un style efficace. En réaction, les aspects textuels défendus par l'approche structuro-cognitive sont apparus comme de meilleurs guides pour la rédaction.

Nous verrons dans la suite de cette section que si certaines de ces critiques ont montré leur pertinence, en particulier en ce qui concerne la simplification de textes, d'autres se sont adoucies suite aux expérimentations conduites au sein du paradigme structuro-cognitif même.

Les apports cognitivistes

Après avoir discouru du procès intenté par les structuro-cognitivistes aux formules de lisibilité, il faut maintenant insister sur les apports indéniables de ce paradigme lorsqu'il est appliqué à l'étude de la complexité textuelle. Il considère l'intérêt que peuvent avoir pour la lisibilité les modèles d'analyse du discours mis au point dans les années 70 et 80 par divers chercheurs en linguistique, rhétorique, psychologie ou même intelligence artificielle. Ces modèles servent essentiellement à utiliser de nouvelles composantes textuelles, de nature structurelle ou cognitive, comme variables indépendantes. Parmi celles-ci, on compte l'organisation du texte (Armbruster, 1984), la cohésion du discours (Clark, 1981; Kintsch, 1979), la structure rhétorique (Meyer, 1982), la charge inférentielle (Clark, 1981; Kemper, 1983; Kintsch et Vipond, 1979), ou encore les représentations mentales préalables des lecteurs.

Le programme structuro-cognitif se révèle également plus ambitieux que les efforts classiques. Parce qu'il fait reposer ses variables indépendantes sur des modèles et des données

expérimentales, il ne se contente pas de répérer des indices textuels **corrélés** principalement avec la difficulté textuelle, mais aspire à identifier des facteurs qui **influencent** l'apprentissage et la mémoire (Binkley, 1988, 98). Autrement dit, il s'agit d'identifier des causes qui expliquent effectivement les divers traitements cognitifs effectués lors de la lecture. Dès lors, cette perspective permet aussi d'apporter une aide à la production, sans tomber dans les travers qui sont reprochés à la simplification de textes basée sur les formules classiques.

Dans cette seconde sous-section, nous envisageons d'abord les travaux portant sur la microstructure, qui se focalisent essentiellement sur les relations de cohésion au sein du discours, mais aussi sur les réseaux cognitifs formés au cours de la lecture (Kemper, 1983; Kintsch, 1979; Kintsch et Miller, 1984; Kintsch et Vipond, 1979). Ensuite, nous présentons quelques études qui prennent en comptes des aspects macrostructurels des textes à analyser (Armbruster, 1984; Meyer, 1982), avant de conclure avec des approches mixtes, qui mêlent facteurs structuro-cognitifs et lexico-syntaxiques.

Kintsch et ses collaborateurs : L'un des chercheurs les plus importants de la tendance structuro-cognitiviste en lisibilité est Walter Kintsch. Connu pour avoir formulé, en collaboration avec le linguiste van Dijk, la première théorie psychologique sur les processus de la compréhension du discours (Kintsch et van Dijk, 1978), il en dérive le premier modèle de lisibilité basé sur des variables structuro-cognitives (Kintsch et Vipond, 1979).

Leur approche repose sur une analyse propositionnelle défendue auparavant par Kintch (1972, cité par Kintsch et Keenan, 1973) : un texte peut être représenté sous la forme d'une liste de propositions, où une proposition correspond à un prédicat accompagné d'un ensemble d'arguments. À partir de cette première hypothèse, Kintsch et Vipond (1979) observent que (1) le temps de lecture est fonction du nombre de propositions dans le texte et (2) les textes qui comprennent un grand nombre d'arguments différents – c'est-à-dire qui introduisent sans cesse de nouveaux concepts – sont plus difficiles à appréhender. Ils postulent que si la répétition d'arguments facilite la lecture, c'est parce qu'elle assure la cohérence du texte. Dans le cas contraire, le lecteur est forcé d'effectuer des inférences pour pallier l'absence de liens explicites entre les propositions.

Par ailleurs, au vu des limites de la mémoire humaine – estimées à plus ou moins sept unités par Miller (1956) et représentées par le symbole s –, il est impossible que le lecteur soit capable de retenir l'ensemble de la structure propositionnelle d'un texte dans sa mémoire de travail. À la place, à chaque cycle de traitement, c'est-à-dire à chaque fois qu'il injecte n propositions dans sa mémoire de travail, il ne conserve qu'un sous-ensemble de la structure totale suite à une étape de réorganisation. Dans ce contexte, lorsqu'une nouvelle proposition ne peut être connectée à la structure contenue dans la mémoire à court terme, le lecteur doit explorer sa mémoire à long terme afin de raccorder cette nouvelle donnée à de l'existant.

Kintsch et Vipond (1979, 348) appellent ce processus réinstallation (*reinstatement*) et estiment que plus il est mis en œuvre pour un texte donné, plus celui-ci peut être considéré comme difficile.

Sur ces bases théoriques, les deux auteurs isolent cinq prédicteurs potentiels de la difficulté :

1. le nombre de propositions par 100 mots ;
2. le nombre moyen d'arguments par proposition ;
3. le nombre d'inférences requises par 100 mots ;
4. le nombre de réinstallations par 100 mots ;
5. le nombre de réorganisations de la structure mémorisée par 100 mots.

Toutefois, ils ne proposent pas de formule dans cet article et sont d'ailleurs conscients que leurs tests, effectués sur quatre passages seulement [47], ne suffisent pas à prouver quoi que ce soit, même s'ils suggèrent que les cinq propriétés textuelles identifiées pourraient constituer de bons prédicteurs de la difficulté.

Kintsch (1979) poursuit ses investigations dans un article postérieur. Il s'appuie toujours sur le même modèle et mesure, chez 120 sujets et sur 20 courts passages, le temps de lecture par proposition remémorée. Bien qu'il dise ne pas vouloir présenter de nouvelle formule de lisibilité, il propose toutefois l'équation suivante (Kintsch, 1979, 10) :

$$\begin{aligned} RD &= 2,83 + 0,48\,RS - 0,69\,WF + 0,51\,PD + 0,23\,INF \\ &\quad + 0,21\,C - 0,10\,ARG \end{aligned} \tag{2.42}$$

où :

RD **:** le nombre de secondes de lecture par proposition remémorée lors d'un test immédiat ;

RS **:** le nombre de réinstallations nécessaires pour un paragraphe ;

WF **:** la fréquence moyenne des mots ;

PDS **:** le nombre de mots du texte divisé par le nombre de propositions ;

INF **:** le nombre d'inférences ;

C **:** le nombre de cycles de traitement ;

47. Il faut noter que, comme la valeur de n varie d'un individu à un autre, et ne peut dépasser s, il existe plusieurs structurations propositionnelles en mémoire pour un même texte. Elles correspondent à différents types de lecteurs. Or, Kintsch et Vipond (1979, 350-351) montrent que le nombre des réinstallations et des réorganisations de la structure diminue inversement à la capacité de la mémoire à court terme. Dès lors, pour paramétriser un texte en fonction d'une population de lecteurs et non d'un seul, les auteurs doivent retenir les valeurs moyennes obtenues, pour les cinq prédicteurs précités, par plusieurs lecteurs chez qui la valeur de s et n fluctue.

ARG **:** le nombre moyen d'arguments par proposition.

Kintsch rapporte un fier $R = 0,97$ et note que les deux meilleurs prédicteurs sont le nombre de réinstallations et la fréquence lexicale. Par contre, le nombre de cycles et le nombre d'arguments n'apportent guère d'informations. On remarque que Kintsch a discrètement réintroduit la fréquence moyenne des mots, une mesure classique qu'il critiquait précédemment et dont il doit bien reconnaître l'efficacité.

Dans un troisième article, Miller et Kintsch (1980) explorent plus précisément ces questions à l'aide d'une vingtaine de textes, testés auprès de 600 sujets. Une procédure de sélection automatique de variables (sélection pas à pas) leur permet d'identifier trois ensembles de facteurs pertinents pour prédire respectivement la vitesse de lecture, le rappel et la lisibilité. Ils combinent cette fois des facteurs traditionnels à ceux qu'ils ont précédemment identifiés. S'ils constatent à nouveau l'importance du nombre de réinstallations et d'inférences sur la lisibilité d'un texte, ils leur faut reconnaître que la fréquence lexicale et la longueur des phrases importent presqu'autant. Ils en viennent ainsi à une vision plus nuancée de la lisibilité classique (Miller et Kintsch, 1980, 348) :

> *the readability of a text is determined by the ways that certain text properties – primarily the arrangement of the propositions in the text base, but also the word frequency and sentence length – interact with the reader processing strategies and resources.*

En conclusion, Kintsch et ses collaborateurs ont proposé un modèle de la compréhension relativement explicite. Ils ont montré son intérêt pour la lisibilité en mettant à jour plusieurs facteurs qui mériteraient d'être considérés au sein d'une étude de plus grande envergure. Toutefois, sa complexité le rend peu pratique, comme y ont insisté Chall et Dale (1995, 99), d'autant qu'il ne décrit pas comment les propositions sont encodées à partir des graphèmes. Il ne prend pas plus en compte la macrostructure du texte, dont Kintsch et van Dijk (1978) ont pourtant souligné l'importance dans le processus de compréhension [48].

Susan Kemper : Proche de Kintsch et van Dijk (1978), Kemper (1983) considère que comprendre un texte revient à organiser temporellement ou causalement une suite d'événements, soit en s'appuyant sur des liens explicites, soit par inférences (en utilisant des connaissances préalables ou des schémas). Dès lors, elle propose une formule de lisibilité qui vise à évaluer la charge inférentielle d'un texte.

Un texte est découpé en un ensemble de propositions (au sens classique du terme) qui correspondent à une action, un état physique ou un état mental, lesquels peuvent être explicites ou

48. Un étudiant de Kintsch, Young, a toutefois proposé dans sa thèse de doctorat (1984, non publiée) une version informatisée du modèle de Miller et Kintsch (1980), qui inclut les aspects macrostructurels.

inférés. Kemper (1983, 399) obtient ainsi six variables qu'elle utilise pour calculer l'équation suivante, dont le $R = 0,76$:

$$IFS = 0,54\,SM - 0,86\,SP + 0,23\,IM + 8,41 \tag{2.43}$$

où :

ILS **:** appelé *Inference Load Score*, il s'agit en fait du niveau scolaire correspondant à un taux de compréhension de 100% à un test de compréhension [49] ;

SM **:** le nombre d'états mentaux explicites ;

SP **:** le nombre d'états physiques explicites ;

IM **:** le nombre d'états mentaux inférés ;

Confronté à cette nouvelle équation, on peut légitimement se demander si elle est meilleure qu'une formule classique. La comparaison de cette formule avec celle de Dale et Chall (1948a), réalisée sur un échantillon de 18 nouveaux extraits, conclut que les deux approches sont équivalentes en termes de pouvoir prédictif. Plus intéressant encore, Kemper (1983) combine ses six facteurs avec ceux qu'ont utilisés Dale et Chall (1948a) et obtient une équation à quatre variables dont le R ne dépasse pas $0,78$. Parmi les quatre facteurs, trois sont cognitifs et le dernier est la longueur moyenne des phrases. Ainsi, l'approche cognitive se heurte à nouveau à la même conclusion : que les variables lexico-syntaxiques sont aussi prédictives que les modèles structuro-cognitifs, nettement plus complexes à utiliser.

Les approches macrostructurelles : Dans une veine macrostructurelle, Meyer (1982) se penche sur la notion de plan, un concept qu'elle définit comme un objectif et une série d'étapes nécessaires pour atteindre cet objectif. Elle s'appuie sur les résultats de Frase (1969) et de Kintsch *et al.* (1977), entre autres, pour avancer : « communication is vastly more efficient (it saves effort) and effective (it gets results) if it follows a topical plan instead of being a miscellaneous sequence of sentences or paragraphs. » (Chall et Dale, 1995, 104). Ainsi, la structure d'un texte peut être envisagée séparément de son contenu. L'approche de Meyer est donc complémentaire par rapport à celles de Kintsch et Kemper.

Meyer (1982) distingue cinq types de plans :

– **antécédent/consequent** : il présente des relations causales du type si/alors ;
– **comparaison** : il compare deux points de vue opposés. On parlera de comparaison **alternative** si la même importance est attribuée aux deux positions ou de comparaison **adversative** dans le cas où l'une est clairement favorisée par rapport à l'autre (ex. discours politique) ;

49. Remarquons que le critère ici utilisé n'est rien d'autre que les leçons de McCall and Crabbs. Étant donné les critiques qui ont été émises par plusieurs chercheurs de l'époque contre ce critère, on pourrait s'étonner de voir Susan Kemper y recourir.

– **descriptif** : il développe un sujet en décrivant certaines de ses parties (ex. article de journaux) ;
– **réponse** : il pose une question avant d'y répondre, par exemple dans le cas d'une séquence problème/solution (ex. article scientifique) ;
– **chronologique** : il présente chronologiquement une séquence d'événements ou d'idées (ex. : traité d'histoire).

Une série d'expériences menées par Meyer (1982) révèlent des faits dignes d'intérêt. Tout d'abord, les bons lecteurs auraient davantage tendance à utiliser le même plan que les auteurs des textes qu'ils lisent. Ce faisant, ils réussissent mieux une tâche de résumé, qu'elle soit réalisée juste après la lecture ou au bout d'une semaine. Ensuite, certains plans semblent plus adaptés que d'autres à certains objectifs communicatifs. Enfin, l'utilisation de connecteurs logiques pour expliciter le plan d'un texte aide les lecteurs moyens à mieux résumer celui-ci. Toutefois, Meyer ne propose pas, à cette époque, de formule de lisibilité basée sur ces observations.

Dans la même lignée que Meyer, se situent les travaux d'Armbruster (1984) et de Calfee et Curley (1984, cités par Chall et Dale, 1995, 108-109). Armbruster synthétise l'influence que diverses caractéristiques textuelles ont sur l'apprentissage. La plus importante serait la cohérence textuelle, qu'elle soit globale – à ce niveau, cinq structures typiques, semblables à celles de Meyer, sont répertoriées – ou locale – à ce niveau, sont envisagés les différents types de liens qui assurent la cohésion des idées. De même, Calfee et Curley insistent sur l'importance des différentes structures textuelles – ils en définissent sept pour leur part –, qu'ils classent de la plus simple (narrative) à la plus complexe (exposition abstraite d'un sujet). Aucune de ces deux approches ne dépasse toutefois la description de facteurs possiblement responsables des difficultés à la lecture.

Enfin, Amiran et Jones (1982) envisagent les effets de trois dimensions déjà étudiées par d'autres cognitivistes sur les difficultés de compréhension. Il s'agit de la structure, de la texture ou charge inférentielle et de la densité de contenu. Leur article présente une sorte de synthèse des efforts structuro-cognitivistes et envisage de nouvelles pistes de recherche à partir de ces résultats.

Approches mixtes : Si l'approche structuro-cognitiviste de la lisibilité a mis au jour nombre de nouveaux facteurs, elle peine à les formaliser en vue d'un traitement quantitatif. C'est pourquoi plusieurs chercheurs du début des années 80 prennent le parti de recourir à un jugement humain guidé, à l'instar du *levelling*. Parmi ceux-ci, Lowe (1979) forge une mesure de lisibilité qui revendique le fait de ne pas être une formule et qu'il appelle ThUS. Elle aide à repérer les Thought Unit Sentences (ThUS) et à évaluer leur difficulté.

Dans le même ordre d'idée, on trouve le système PHrase ANalysis (PHAN) de Clark (1981)

qui reprend et simplifie le modèle d'analyse propositionnelle de Kintsch et van Dijk (1978). Cela revient à découper un texte en syntagmes et à compter les trois variables suivantes :

– le nombre de syntagmes qui séparent un pronom de son référent ou deux informations connectées ;
– le nombre d'inférences requises ;
– le nombre de concepts complexes en contexte.

Ces trois variables ne sont toutefois pas combinées sous la forme d'une équation. Pour Clark (1981), elles ne doivent servir que comme éléments de comparaison entre textes. Sa technique ne permet donc pas d'attribuer un score absolu de difficulté.

Plus complexe est la liste de contrôle de Irwin et Davis (1980). Les auteurs proposent une liste de 45 questions qu'un professeur devrait se poser pour évaluer la difficulté d'un texte. Ces questions explorent divers aspects du texte – son niveau de compréhension, le taux d'apprentissage qu'il autorise, sa capacité à réactiver des connaissances antérieures ou encore son taux d'intérêt – et incorporent divers facteurs, négligés par les formules classiques, et sélectionnés sur la base de recherches en psychologie cognitive.

Autres formules

À côté de ces divers efforts structuro-cognitivistes, paraissent quelques travaux de tendance plus classique, qui n'hésitent pas à pratiquer l'autocritique. Maxwell (1978) suggère qu'on a peut-être été trop loin dans l'utilisation des formules de lisibilité et que l'on a paru oublier qu'il ne s'agit que d'outils imparfaits. Quant à Harris et Jacobson (1979), ils appellent à dépasser les variables traditionnelles, précisant qu'il s'agit du seul moyen d'effectuer enfin des avancées significatives en lisibilité.

Leurs propositions ne sont toutefois pas suivies par la principale réalisation de l'époque : un système d'évaluation du niveau maximal de compétence en lecture d'un étudiant, qui est développé par le College Entrance Examination Board (1980, cité par Kibby, 1981). Ce test, appelé Degrees of Reading Power (DRP), consiste en une série de questions portant sur des passages dont la difficulté croît au fur et à mesure, jusqu'à ce que la personne évaluée ne soit plus capable de répondre correctement. On estime alors qu'il a atteint son niveau de compétence maximale en lecture, lequel est évalué sur une échelle absolue, exprimée en fonction du texte le plus compliqué qu'un étudiant du même niveau peut comprendre.

Pour obtenir ces valeurs absolues, les concepteurs de DRP ont utilisé le résultat (Res) de

l'une des formules de Bormuth (1969a) [50], modifié de la sorte (Kibby, 1981, 419) :

$$DRP = (1 - Res) \text{ x } 100 \tag{2.44}$$

où Res correspond au score moyen qui serait obtenu à un test de closure sur ce texte. Il s'agit donc à première vue plutôt d'une application originale de la lisibilité au domaine du testing que d'une formule véritablement nouvelle. Toutefois, comme les concepteurs de DRP ont évalué un grand nombre de textes utilisés dans l'éducation à l'aide de cette formule, il est devenu possible de situer sur la même échelle un lecteur et un texte [51].

Cette approche a toutefois fait l'objet de critiques. Kibby (1981) lui reproche, entre autres, l'emploi d'une formule de lisibilité de type classique pour étalonner les textes qui servent ensuite à tester les étudiants. Il montre que certains textes, considérés comme faciles, ont posé problème de manière générale et attribue ce phénomène aux limitations de la formule de Bormuth (1969a), qui ne prend pas en compte les connaissances préalables, la motivation et l'intérêt des lecteurs. Quant à Carver (1985), il met en évidence des incohérences dans l'échelle utilisée pour situer les élèves et les textes. Ainsi, si le score DRP moyen des livres simplifiés (*basal reader* en anglais) de la première année primaire est 41, le score de la moitié des quatrièmes années primaires est estimé à 34,5. On a du mal à croire qu'un livre destiné à un élève de première année soit trop complexe pour un élève de quatrième.

À l'exception des DRP, nous n'avons pas identifié d'autres travaux notables dans la veine classique. DuBay (2004) précise cependant que le début des années 80 voit l'apparition de divers programmes informatisés qui proposent non seulement des formules de lisibilité, mais aussi d'autres types d'assistances à l'écriture telles que le repérage des mots complexes, des phrases longues, etc. Le plus célèbre d'entre eux serait le *Writer's Workbench* développé par les Bell Laboratories (Macdonald *et al.*, 1982).

La lisibilité francophone

Les travaux francophones de cette période sont, pour une fois, en phase avec les recherches anglophones. L'approche structuro-cognitive des textes fleurit en effet des deux côtés de l'Atlantique et stimule l'apparition d'une série d'articles en français sur la lisibilité. Ceux-ci reproduisent les critiques présentées précédemment. Pour Foucambert (1983), les formules de lisibilité oublient les particularités du lecteur et tendent à confondre les signes et les causes de la difficulté. Plus précis dans sa critique, Benoît (1986b) estime que les critères utilisés pour la conception des formules peuvent aujourd'hui être considérés comme inva-

50. Les détails de la formule sont rapportées par Klare (1984, 697). Elle utilise comme variables le nombre de lettres par mots, le pourcentage de mots sur la liste de 3000 mots de Dale et le nombre de mots par phrase.

51. Notons que le test est toujours utilisé aujourd'hui et qu'il existe un site internet qui lui est dédié.

lides et que les variables du lecteur et les aspects qualitatifs du langage sont oubliés, ce qui s'explique notamment par le manque d'assises théoriques des travaux en lisibilité.

Dans la première moitié des années 80, les réalisations concrètes ne foisonnent pas. Fortier (1982, cité par Boyer, 1992, 5) adapte la formule de Henry (1975) au contexte québécois et en fournit une version informatisée. À l'opposé, Timbal-Duclaux (1984) s'intéresse à des facteurs structuro-cognitifs, uniquement analysables via une procédure manuelle. Pour simplifier celle-ci, il imagine une procédure en quatre étapes qui se limite à surligner, en vert ou en rouge, certaines caractéristiques désirables ou indésirables du texte à évaluer. Si la proportion de rouge est trop importante, il convient alors de réviser le texte. Les caractéristiques prises en considération sont la subordination des détails, la variété des structures, la quantité de mots abstraits et le ratio des verbes au passif. Il ne s'agit toutefois que de conseils théoriques : l'auteur ne fournit aucun résultat expérimental qui justifierait la méthodologie et les variables utilisées.

Dans un article postérieur, Timbal-Duclaux (1989) revisite et automatise cette idée. Il implémente DUMARSAIS, un programme en LISP qui effectue l'analyse qualitative de textes selon quatre dimensions, chacune identifiée par une couleur différente : en rouge, les concepts, c'est-à-dire les idées que veut faire passer le texte ; en bleu, les idées admises, qui servent au lecteur à rattacher les concepts à ce qu'il connaît ; en vert, les procédés stylistiques, qui occupent une fonction « phatique » et en noir, les indices de la structure. Timbal-Duclaux estime que bien écrire revient à doser savamment ces quatre couleurs. Trop de rouge et le texte semblera trop difficile ; trop peu, et il paraîtra banal et fade. De même, un texte où domine le vert risque d'être perçu comme « démagogique » (*sic*.). Ainsi, le logiciel de Timbal-Duclaux (1989) est plus un outil d'aide à la rédaction qu'une formule de lisibilité.

À la fin de la décennie, des références aux travaux de Kintsch et van Dijk (1978) commencent à apparaître dans les publications francophones en lisibilité. Ainsi, Racle (1988) décrit le concept d'Unité Sémantique Abstraite Élémentaire (USAÉ), qui correspond aux propositions de Kintsch. Il circonscrit également les diverses dimensions du processus de lecture, qui sont autant de facteurs potentiels de lisibilité, mais il n'en expérimente aucun. Comme trop souvent dans les articles francophones de l'époque, le discours se limite à des considérations théoriques. Une autre présentation de ces aspects cognitifs se retrouve dans l'excellente synthèse de Gromer et Weiss (1990), qui les met en balance avec les formules classiques.

Notons enfin la parution d'articles qui appliquent des formules de lisibilité à divers types de texte. Ainsi, Tremblay (1984) applique les formules de Flesch (1948) et de Gunning (1952) à un corpus de journaux francophones montréalais, tandis que Delory (1989) analyse la difficulté des examens interdiocésains belges à l'aide des formules de Flesch (1948) et

d'Henry (1975).

Dans l'ensemble, le domaine reste toutefois assez anonyme. Timbal-Duclaux (1985, 29) considère que cette différence de succès dans les mondes anglophone et francophone s'expliquerait par des raisons culturelles :

> *Les formules de lisibilité issues du monde américain véhiculent la culture anglo-saxonne. Si nous, les héritiers des Latins, les avons toujours majoritairement ignorées ou refusées, ce n'est pas un hasard. C'est parce qu'elles portent en elles une « vision du monde » qui n'est pas la nôtre.*

Conclusion

Pour conclure, revenons sur la perspective structuro-cognitiviste, qui domine largement cette période et a profondément ébranlé le domaine de la lisibilité, tant par ses critiques que par les perspectives qu'elle offre. Au terme d'une décennie d'explorations, elle offre toutefois un bilan mitigé que Boyer (1992, 11) résume ainsi : « il est actuellement impossible de remplacer les anciennes formules par une nouvelle qui serait plus valide ». Autrement dit, les efforts structuro-cognitivistes n'ont pas abouti à un modèle à la fois efficace et manipulable, et cela s'explique par plusieurs faits.

Tout d'abord, contrairement aux espoirs suscités par ces mêmes articles qui fustigeaient les formules classiques et leurs variables traditionnelles, les nouvelles dimensions textuelles découvertes ne se révèlent pas de meilleurs prédicteurs de la difficulté que les indices lexico-syntaxiques. D'ailleurs, les critiques des structuro-cognitivistes tendent à s'atténuer tandis que ceux-ci se rendent comptent que (1) leurs propres formules ne sont pas plus efficaces que les approches classiques alors qu'elles sont nettement plus complexes à mettre en œuvre ; (2) la combinaison de variables classiques et structuro-cognitives n'améliore guère les performances, ce qui laisse supposer que ces deux types d'informations sont fortement corrélées entre elles – comme l'avait déjà noté Bormuth (1966). Par contre, les chercheurs espèrent tout de même que ces nouveaux facteurs découverts entretiennent un véritable lien de causalité avec la difficulté et soient plus utiles pour la production.

Ensuite, les modèles structuro-cognitivistes feront eux-mêmes l'objet de critiques. D'une part, parce qu'ils n'envisagent qu'une partie du processus de compréhension, que ce soit à travers la cohésion microstructurelle, la cohérence globale, ou, au mieux, de quelques facteurs cognitifs. Par exemple, le modèle de Miller et Kintsch (1980) ne prend pas en compte la motivation du lecteur, ses connaissances préalables, ou encore la situation de lecture. D'autre part, la validité des prédictions effectuées par ces nouvelles formules semble tout autant sujette à caution que celle des formules classiques. C'est ce que révèle l'étude de

Trollinger et Kaestle (1986), lesquels ont utilisé la formule de Kemper (1983) pour évaluer la difficulté de journaux américains des années 20. D'après cette formule, la difficulté du *Post* serait à un niveau de lecture équivalent à la seconde primaire, ce qui semble malaisé à croire. Chall et Dale (1995, 101) considèrent que cette formule est plus efficace que celle de Flesch – aussi utilisée dans l'étude de Trollinger et Kaestle (1986) – pour évaluer la difficulté relative de textes, mais qu'elle l'est moins pour obtenir une valeur absolue, ce qui est nécessaire lorsqu'on veut associer un texte à un niveau d'éducation donné.

Toutefois, l'intérêt du mouvement structuro-cognitiviste apparaît lorsqu'on prend de la hauteur et cesse de l'envisager en confrontation avec la lisibilité classique. En réalité, ces deux points de vue sont complémentaires et chacun d'eux propose sur les techniques d'évaluation de la difficulté textuelle un discours qui possède ses qualités et ses faiblesses. L'approche classique peut s'appuyer sur un legs important de travaux et d'observations et vise avant tout à l'efficacité pratique. Par contre, ce que l'on appelle aussi la « nouvelle lisibilité » pose un discours plus théorique que pratique, mais dont la pertinence est justifiée par les résultats d'expériences psycholinguistiques. Mais, plus que toute autre chose, la principale différence entre ces deux démarches vient de ce qu'elles n'utilisent pas le même type de variables, et partant ne s'attachent pas au même instant du processus de lecture, ainsi que l'ont souligné Amiran et Jones (1982) ou Clark (1981) entre autres. Les variables traditionnelles se concentrent sur le décodage, tandis que les variables structuro-cognitives se focalisent sur les processus de compréhension (voir Chapitre 1).

La pertinence de cette distinction est renforcée par une observation répétée par plusieurs chercheurs de l'époque. Il semblerait que les mesures cognitivo-structurales soient plus discriminantes pour du matériel de haut niveau, c'est-à-dire lorsque le contenu importe plus que la langue, tandis que les formules classiques se comportent mieux pour les textes plus faciles (Armbruster, 1984; Carrell, 1987). Un résultat récurrent que Chall et Dale (1995, 108) expliquent ainsi :

> *For children or adults who may not have reached the linguistic and knowledge base used in most content area textes, word and sentence factors tend to be excellent predictors of difficulty. For readers who may have a considerable vocabulary and knowledge base, the structural measures of Meyer and Armbruster as well as the cognitive measures of Kintsch and Kemper may be better predictors.*

En conclusion, nous nous accordons avec Chall et Dale (1995, 112) à dire que la recherche en lisibilité doit viser une synthèse entre les facteurs classiques et structuro-cognitifs, ce qui ne va être que trop rarement le cas, comme nous le verrons par la suite.

2.2.7 L'après-cognitivisme : le pragmatisme désenchanté (1990 à 2003)

Introduction

Nous avons intitulé la période suivante, l'après-cognitivisme ou le pragmatisme désenchanté. En effet, après l'engouement structuro-cognitiviste, qui perdure jusqu'au début des années 90 dans le monde anglo-saxon et quelques années de plus en France, l'enthousiasme va peu à peu retomber suite à l'absence de modèles réellement utilisables et pratiques pour le milieu éducatif. Ce changement est observé par Chall et Dale (1995, 112-113) qui recensent dans ERIC [52] les publications en lisibilité entre 1984 et 1994. Avant 1990, les travaux sont majoritairement d'inspiration cognitiviste. Par après, sur les onze entrées détectées, une seule suit encore ce paradigme et quatre autres mélangent des variables traditionnelles et structuro-cognitives. On peut donc bien parler d'une sorte de retour à l'approche classique, même si la parenthèse cognitive laisse – heureusement ! – des traces.

De manière générale, la retombée d'enthousiasme se marque aussi par une diminution sensible des publications dans le domaine. Nous présentons dans cette section les plus notables d'entre eux. Une fois n'est pas coutume, nous commençons par l'une des rares entreprises à évoluer dans le monde de la lisibilité : MetaMetrics et son *The Lexile Framework*. Suit ensuite la description de travaux d'inspiration classique : Chall et Dale (1995), Fry (1989, 1990) et Paul (2003). Nous terminons par deux études structuro-cognitivstes dues à Kemper *et al.* (1993) et Meyer *et al.* (1993).

The Lexile Framework

MetaMetrics est une société qui existe encore aujourd'hui et qui vend depuis plus de vingt ans une mesure de lisibilité appelée *The Lexile Framework*, développée par Stenner *et al.* (1988). À l'instar des DRP, ces travaux cherchent à situer des textes et les lecteurs sur une même échelle de difficulté, mais ils veulent également développer un instrument de mesure objectif, c'est-à-dire qui fournit des mesures indépendantes de l'instrument utilisé et du contexte dans lequel celui-ci est employé. Pour ce faire, Stenner (1996) recourt au modèle de Rasche (1961) et définit l'échelle Lexile, qui s'échelonne de 0 à 2000.

Au final, si ces travaux résultent sur un instrument de mesure de la compréhension qui se veut objectif, l'équation prédisant cette valeur reste très classique. Elle utilise le logarithme de la longueur moyenne des phrases ($LMSL$) et la moyenne du logarithme des fréquences

52. La base ERIC (Education Resources Information Center) est une base bibliographique spécialisée dans les parutions scientifiques dans le domaine de l'éducation.

lexicales ($MLWF$) (Stenner et Burdick, 1997, 9-10) :

$$\text{Lexile text measure} = (9,82247\, LMSL - 2,14634\, MLWF + 3,3) * 180 \qquad (2.45)$$

Une version informatisée de la formule a rapidement été rendue disponible. Aujourd'hui, il existe aussi un site [53] dédié au *Lexile Framework*. Il permet notamment à un lecteur connaissant son niveau en Lexile de rechercher des textes de difficulté adaptée.

Formule révisée de Dale and Chall

Une autre entreprise majeure de cette période est la révision de la célèbre formule de Dale et Chall (1948a) par les auteurs eux-mêmes (Chall et Dale, 1995). Elle repose toujours sur les deux mêmes prédicteurs : la difficulté lexicale et la longueur des phrases, mais plus sur le même critère. Cette fois, c'est le test de closure qui a été retenu. Les deux auteurs ont utilisé les corpus de Bormuth (1971, cité par Chall et Dale, 1995, 60), de Miller et Coleman (1967), de Caylor *et al.* (1973) et de MacGinitie et Tretiak (1971) pour tester l'efficacité de leurs variables. Finalement, c'est la formule entraînée à partir des 32 passages de Bormuth qui a été retenue (Chall et Dale, 1995, 66) :

$$\text{Dale-Chall cloze} = 64 - 0,95\, X_1 - 0,69\, X_2 \qquad (2.46)$$

dans laquelle :

Dale-Chall cloze : pourcentage moyen de bonnes réponses à un test de closure sur ce texte ;

X_1 **:** le pourcentage de mots absents de la liste revisée de Dale, qui comprend les 3000 mots connus par 80% des enfants de quatrième primaire ;

X_2 **:** longueur moyenne des phrases en nombre de mots.

Cette équation présente un taux de corrélation multiple avec le critère de $0,92$, ce qui très élevé. Toutefois, il n'est pas à exclure que ce gain par rapport à la première formule provienne davantage de la nature du critère utilisé que de l'amélioration de la formule, puisque ce corpus de Bormuth est connu pour produire des R élevés. Par ailleurs, nous sommes méfiant vis-à-vis du fait que la formule retenue l'ait été via la maximisation du R en fonction des données, sans recours à un échantillon de validation. En effet, rien ne permet d'affirmer que cette supériorité soit généralisable. Une validation des neuf formules sur un même corpus de test aurait sans doute permis d'être plus assertif sur cette question, pour autant que les résultats des différentes formules aient été exprimés selon une échelle commune.

53. L'adresse était la suivante au 01-12-2010 : http ://lexile.com/findabook/ .

Il est possible que Chall elle-même n'ait pas été pleinement convaincue par sa nouvelle formule, puisqu'elle publie l'année suivante une autre méthode d'évaluation des textes, qui utilise le *levelling* (Chall *et al.*, 1996). Cette seconde approche, plus globale, n'est pas sujette aux limitations des formules classiques, notamment en ce qui concerne la prise en compte des aspects syntaxiques et sémantiques. À l'instar de Carver (1975), les auteurs estiment qu'il est possible d'estimer la complexité d'un texte en le comparant à des passages de références, mais ils vont plus loin et définissent six références qui divergent par le sujet des extraits. Dans ce même paradigme du *levelling*, citons également les travaux de Clay (1991), Fountas et Pinnell (1999) et Weaver (2000).

Nouveaux travaux de Fry

Edward Fry, après avoir réaffirmé la validité des formules classiques pour peu que l'on évite certaines mauvaises pratiques (Fry, 1989), en propose une qu'il estime apte à traiter de courts passages (Fry, 1990). La difficulté d'une telle tâche réside dans le fait qu'un court extrait ne contient guère de données, ce qui rend une estimation fournie par une formule peu fiable.

L'équation de Fry (1990, 595) est très simple :

$$\text{Readability} = \frac{\text{Word Difficulty} + \text{Sentence Difficulty}}{2} \tag{2.47}$$

Pour obtenir la difficulté lexicale, l'analyste n'a qu'à sélectionner trois mots qu'il considère essentiels pour comprendre le passage, à rechercher leur difficulté (exprimée en termes de niveaux scolaires) dans le *Living Word Vocabulary* de Dale et O'Rourke (1981) et à calculer la moyenne de ces niveaux. Quant à la difficulté phrastique, elle se calcule de manière semblable. Pour chaque phrase, on compte le nombre de mots qu'elle contient, lequel est ensuite transformé en un niveau scolaire à l'aide d'une table fournie par Fry [54]. La difficulté phrastique correspond à la moyenne des *grades* de chaque phrase.

La formule ATOS

Des chercheurs du School Renaissance Institute et de Touchstone Applied Science Associates (TASA) ont produit la formule Advantage-TASA Open Standard (ATOS) (Paul, 2003). Leur objectif est de créer une formule gratuite et simple à utiliser. Pour ce faire, ils recourent aux données de plus de 30 000 étudiants ayant lu et été testés sur près de 950 000 livres. Bien que plus de 100 variables différentes aient été testées, leur formule (décrite dans Paul, 2003, 67-68) utilise trois prédicteurs tout à fait classiques : la longueur des phrases, le niveau moyen

54. Cette table a été obtenue à partir de son graphe (Fry, 1977) et associe un niveau scolaire (ou *grade*) à la difficulté médiane des phrases de ce niveau.

du vocabulaire (en *grades*) et la longueur des mots. Ils obtiennent de très bonnes corrélations entre ces trois variables et leur critère, à savoir des résultats de tests de compréhension. Cette formule, dont il existe une version automatisée, est aussi la première à analyser des livres en entier et non à partir de quelques échantillons.

Formules structuro-cognitivistes

Malgré ce retour vers des variables plus traditionnelles, certains chercheurs de la mouvance structuro-cognitivistes poursuivent et élargissent leurs recherches. Ainsi, Kemper *et al.* (1993) et Meyer *et al.* (1993) abordent une thématique nouvelle : la lisibilité de textes pour un public âgé (à partir de 60 ans). Ces auteurs suggèrent que cette catégorie de personnes éprouve davantage de difficulté à lire. Avec l'âge, les capacités de leur mémoire de travail déclinent et certains mécanismes inhibiteurs faiblissent, permettant l'irruption de pensées hors de propos durant la lecture.

Pour Kemper *et al.* (1993), ces deux phénomènes rendent plus ardu le traitement de textes comprenant des phrases dont la structure syntaxique est complexe et la densité propositionnelle plus importante. Ce sont en effet ces deux facteurs qui ressortent d'une étude où le niveau de compréhension, la vitesse de lecture et l'intérêt déclaré sont corrélés à diverses variables mesurant la difficulté du contenu, la longueur des phrases, la quantité et le type de structures enchâssées. L'article ne présente toutefois pas de formule précise, alors qu'une analyse de régression a pourtant été effectuée.

De leur côté, Meyer *et al.* (1993) s'intéressent à l'accessibilité des documents quotidiens pour les personnes âgées. Toutefois, les auteurs se voulant méthodologiquement corrects, ils ne définissent pas leur formule comme une mesure de la lisibilité de ce type de documents. S'inspirant de la recherche de Kirsch et Mosenthal (1990), ils mesurent plutôt la complexité des items dans les tests de compréhension portant sur ce type de textes. Dès lors, parmi les facteurs envisagés, on trouve la position de la réponse dans le texte, l'efficacité des distracteurs ou encore la proximité entre la réponse et l'item. Ces variables sont incluses dans une équation de régression dont les coefficients sont estimés à partir des résultats de 483 sujets à 62 items. Au-delà de l'intérêt de cette étude en ce qui concerne la prédiction de la complexité d'items d'un test de compréhension, les auteurs estiment que leurs résultats révèlent la capacité de facteurs discursifs à prédire la difficulté de compréhension de tâches de lecture quotidiennes.

La lisibilité francophone

Clôturons cette période par la présentation des travaux les plus récents sur le français L1. À l'image de la situation anglophone, les recherches de cette époque en français sont rares, comme le rappelle Mesnager (2002). Toutefois, la critique s'atténue également. Ainsi, Labasse (1999, 86) affirme : « ces formules, et en particulier les plus simples, demeurent, malgré leurs imperfections, d'une réelle utilité pour les praticiens et les formateurs ». On trouve également quelques synthèses intéressantes qui réconcilient les points de vue classique et cognitif (Beaudet, 2001; Boyer, 1992; Gélinas Chebat *et al.*, 1993).

L'auteur le plus actif de cette période est Jean Mesnager. Dès 1986, il lance un projet de formule pour les enfants en collaboration avec Jean-Pierre Benoît. Partant de l'idée bien connue qu'un jeune lecteur confronté à des textes trop ardus pour lui en vient rapidement à abandonner la lecture, Mesnager (1986a) remarque que de nombreux livres pour enfants abusent de vocables complexes, de phrases longues et rechignent à répéter les mêmes mots alors que la répétition améliore la cohésion du texte. Il se propose d'associer harmonieusement lecteurs et textes. Dans la seconde partie de l'article (Mesnager, 1986b), il envisage divers facteurs de lisibilité, aussi bien classiques (nombre de lettres par mots, longueur des phrases) que discursifs (cohésion, les repères du récit), sans proposer encore de formule. Pour sa part, Benoît (1986a,c) rapporte diverses applications des formules de Henry (1975) à des textes informatifs destinés à des élèves du secondaire.

Il faut attendre trois ans pour que Mesnager (1989) publie sa propre formule. En effet, il considère que la formule d'Henry (1975) est trop peu commode et que la liste lexicale utilisée (Gougenheim *et al.*, 1964) commence à dater. Il cherche d'abord à améliorer la formule de Flesch (1948), « pour essayer d'en tirer quelque chose d'un peu plus fiable que le bricolage de De Landsheere (sauf le respect à lui dû) » (Mesnager, 1989, 20). Toutefois, il observe, au terme de deux expérimentations, que le nombre de syllabes pour 100 mots n'est pas un aussi bon prédicteur que le nombre d'absents de la liste de Catach (1985)[55]. Sa formule finale s'inspire donc plutôt de Dale et Chall (1948a) :

$$\text{Difficulté} = \frac{2}{3}\,AC + \frac{1}{3}P \tag{2.48}$$

dans laquelle :

Difficulté : est un résultat compris entre 6 (très facile) et 25 (très ardu) ;

AC **:** le pourcentage de mots absents de la liste de Catach (1985) ;

P **:** longueur moyenne des phrases en nombre de mots.

55. Il s'agit d'une liste de 3000 mots, dont il convient d'enseigner l'orthographe en priorité aux élèves de primaire.

On notera toutefois que méthodologie n'est pas aussi stricte que celle habituellement utilisée en lisibilité. Les coefficients de l'équation sont estimés au jugé, sans recourir à un corpus d'entraînement. De même, la validation n'a été réalisée que sur huit textes et il est donc délicat d'associer un score de la formule avec un niveau scolaire déterminé.

De l'autre côté de l'Atlantique, Daoust *et al.* (1996) développent SATO-CALIBRAGE, un outil informatique d'assistance pour le choix et la rédaction de textes qui peut attribuer à un texte un indice de difficulté et le situer sur un continuum s'étendant de la première primaire à la cinquième secondaire (système canadien). Ils rassemblent un corpus de 679 textes, classés en fonction du niveau scolaire auquel ils sont destinés et repèrent automatiquement plus de 120 variables à l'aide du logiciel SATO [56]. Parmi celles-ci, une analyse de régression multiple hiérarchique leur permet de formuler une équation combinant pas moins de 14 variables (cf. Daoust *et al.*, 1996, 223) et qui atteint un R de $0,86$.

Cet outil intéressant annonce une nouvelle évolution du domaine, que nous aborderons dans la section suivante. Cette évolution s'explique par l'application de techniques issues de la linguistique computationnelle pour repérer automatiquement des facteurs plus complexes. Ainsi, le pourcentage de verbes conjugués et de locutions grammaticales complexes sont, par exemple, considérés dans cette étude. En conséquence, la formule retenue comporte nettement plus de variables que les standards habituels. Il est d'ailleurs possible qu'elle soit sujette à un problème de surentraînement, étant donné que les auteurs ne rapportent pas de résultats en validation croisée.

Conclusion

Durant cette période intermédiaire, la lisibilité semble à nouveau engourdie. Contronté à l'échec du paradigme structuro-cognitiviste d'arriver à une formule efficace et simple d'application, les chercheurs semblent hésiter sur la marche à suivre. Beaucoup reviennent à une approche classique, soit en défendant son intérêt à l'aide d'arguments théoriques (Chall et Dale, 1995; Mesnager, 1989), soit en l'appliquant à grande échelle (Paul, 2003; Stenner, 1996). On peut cependant douter qu'ils apportent grand-chose de neuf au domaine. Parallèlement, d'autres s'obstinent dans le paradigme structuro-cognitif (Kemper *et al.*, 1993; Meyer *et al.*, 1993), sans plus parvenir à proposer des outils aisément utilisables.

Entre la simplicité d'emploi des ses outils et la volonté d'exhaustivité dans la prise en compte des facteurs de la difficulté, le chercheur en lisibilité est confronté à un sérieux dilemme. Pourtant, depuis le Québec, Daoust *et al.* (1996) suggèrent une voie à suivre : et si la synthèse entre approche classique et approche structuro-cognitive passait par le TAL ? C'est

56. Ce logiciel est disponible en ligne à l'adresse suivante : `http://www.ling.uqam.ca/sato/`.

l'hypothèse que nous développons dans la section suivante.

2.2.8 La lisibilité computationelle (2004 à nos jours)

Introduction

Nous avons vu que l'arrivée de l'informatique a modifié les pratiques en lisibilité dès les années 60 (Danielson et Bryan, 1963). L'un des défis de l'automatisation consiste à écrire des programmes aptes à compter fiablement les variables linguistiques utilisées. Parmi celles-ci, c'est le nombre de syllabes par mots – facteur employé par Flesch (1948) – qui a fait l'objet des premiers efforts. Ainsi, Fang (1968) implémente une routine pour laquelle une syllabe équivaut à une voyelle, moyennant certaines exceptions. Klare *et al.* (1969) proposent d'autres règles, sans parvenir non plus à atteindre les 100% d'exactitude [57].

De manière générale, malgré les efforts consentis, l'automatisation entraîne une simplification des réalités linguistiques. Ainsi, pour McCallum et Peterson (1982, 47), un mot est une séquence de lettres séparées par des blancs ou des ponctuations ; une phrase est délimitée par des points, des points d'exclamation ou d'interrogation et les abréviations sont souvent mal gérées. Coke et Rothkopf (1970) en viennent à défendre le simple décompte de voyelles comme moyen d'estimation du nombre de syllabes.

Parallèlement à ces efforts, d'autres chercheurs, dont Smith (1961) le premier, réfléchissent à des formules qui n'emploieraient que des variables manipulables par une machine. Par exemple, c'est à partir de ce moment que le nombre de lettres par mot va être préféré au nombre de syllabes. Cette seconde stratégie, si elle résout – en partie – les problèmes de comptage des formes linguistiques, présente un revers important. Elle conduit nombre de formules à n'envisager que certains types de facteurs, dont la nature trop superficielle a rapidement été critiquée par les structuro-cognitivistes.

Malheureusement, si la « nouvelle lisibilité » recentre l'attention sur des facteurs qui entretiennent plus probablement une relation causale avec la compréhension, elle ne fournit guère de moyens pratiques de paramétriser automatiquement ces variables dans des textes, comme nous l'avons vu. Ce qui cause son déclin. Toutefois, divers chercheurs se refusent à abandonner ces pistes fructueuses et font appel à diverses techniques issues du TAL afin d'analyser plus finement des textes avec un ordinateur. Ces quelques pionniers (Daoust *et al.*, 1996; Foltz *et al.*, 1998) vont faire florès et, à partir de 2005, le nombre d'études alliant la lisibilité et le TAL explose.

57. Signalons que leur programme, doté de nombreuses options, automatise également les formules de Gunning (1952), de Farr *et al.* (1951) et le Fog Count de McElroy.

En sus des techniques de TAL, la lisibilité va également évoluer considérablement sous l'influence d'une seconde discipline : l'intelligence artificielle et ses nombreux algorithmes statistiques (dont certains sont résumés au Chapitre 5). Ce sont Si et Callan (2001) qui, les premiers, vont envisager ce mariage entre les deux domaines, lequel a rencontré un tel succès qu'il n'est aujourd'hui plus évident de développer une formule de lisibilité à l'aide d'une régression linéaire multiple.

Les pionniers

Avec Daoust *et al.* (1996), le premier à relancer le domaine de la lisibilité en manque d'innovations est, une fois de plus, Walter Kintsch, cette fois en compagnie de Foltz et Landauer (Foltz *et al.*, 1998). Les trois auteurs présentent une nouvelle mesure de la cohérence des textes, basée sur une approche automatisée découverte quelques années avant : l'analyse sémantique latente (LSA) (Deerwester *et al.*, 1990; Landauer *et al.*, 1998). Ils parviennent à montrer que cette mesure est corrélée avec la compréhension.

Plus précisément, les auteurs commencent par définir un espace vectoriel sur la base d'un grand nombre de documents portant sur des sujets similaires à ceux des textes à analyser. Dès lors, chacune des phrases du texte peut être représentée dans cet espace sous la forme d'un vecteur et la cohérence entre deux phrases adjacentes correspond au cosinus de leurs vecteurs. La cohérence globale d'un texte équivaut alors à la moyenne des cosinus pour chaque paire de phrases adjacentes. Foltz *et al.* (1998) effectuent des expérimentations sur deux corpus différents et montrent que la cohérence textuelle estimée par la LSA est corrélée avec le degré de compréhension de ces mêmes textes par des lecteurs faibles.

Cette démarche, pour intéressante qu'elle soit, fait moins d'émules qu'une seconde approche, issue des sciences informatiques et initiée par Si et Callan (2001) de la *Carnegie Mellon University*. Celle-ci doit son succès à une reformulation de la problématique de la lisibilité sous les traits d'un problème de classification, ce qui autorise le recours à des techniques statistiques discrètes. Pour Si et Callan (2001, 575), évaluer la difficulté d'un document d_i revient à lui attribuer la classe [58] g qui maximise la probabilité suivante :

$$P_c(g \mid d_i) = \lambda * P_a(g \mid d_i) + (1 - \lambda) * P_b(g \mid d_i) \tag{2.49}$$

P_c est en fait la combinaison linéaire de deux autres probabilités : P_a, qui est obtenue à l'aide d'un modèle de langue, et P_b, qui modélise la distribution de la longueur des phrases sous la forme de gaussiennes. Le modèle de langue P_a vise à mieux prendre en compte le

58. Notons que les auteurs se limitent à un problème à trois classes : k-2 (du jardin d'enfants à la seconde primaire) ; 3-5 (de la 3e à la 5e primaire) et 6-8 (de la 6e primaire à la 2e secondaire).

contenu des textes et il est approximé à l'aide d'un unigramme. Calculer P_a revient donc à sommer $\log P(w_i \mid g)$, soit le logarithme de la probabilité de chaque mot étant donné le niveau. L'hypothèse sous-jacente est qu'il existe des différences notables dans l'utilisation des mots selon les niveaux, comme le confirment par la suite Collins-Thompson et Callan (2004a).

Cette formule ne ciblant que des textes scientifiques, les auteurs entraînent le modèle unigramme sur des syllabus en sciences et en mathématiques et obtiennent la valeur de λ à l'aide d'une procédure Expectation-Maximization (EM) appliquée à un corpus de 91 pages internet dont le niveau a été annoté par leurs concepteurs. La valeur optimale retenue est $0,91$, ce qui signifie que le modèle unigramme est, de loin, la source d'information principale. Par ailleurs, la validité de cette nouvelle approche est confirmée par sa supériorité, sur ce corpus, vis-à-vis de la formule de Flesch (1948). On pourrait toutefois objecter que le critère utilisé pour mesurer le taux de compréhension semble moins précis que dans d'autres études récentes. Nous envisagerons cette question dans le Chapitre 4.

De ces deux études fondatrices naissent trois orientations. La première, qui domine jusqu'en 2008, va approfondir la démarche de Si et Callan (2001). Si elle se limite à des facteurs lexicaux et syntaxiques assez traditionnels, abordés selon une optique computationnelle, elle renouvelle les manières de les combiner en empruntant diverses techniques statistiques au domaine de l'apprentissage automatique. Les travaux les plus marquants de cette mouvance sont dus à Collins-Thompson et Callan (2004a), Schwarm et Ostendorf (2005) et Heilman *et al.* (2008). Profitant des acquis de ces travaux, une seconde tendance se dégage à partir de 2007. Représentée par Crossley *et al.* (2007) ou Pitler et Nenkova (2008), elle vise à développer des formules de lisibilité automatisée qui soient enfin capables de prendre en compte les aspects structuro-cognitifs et de les combiner avec les variables traditionnelles. Enfin, certains chercheurs, tels que Pitler et Nenkova (2008) et Tanaka-Ishii *et al.* (2010), partent du constat qu'il est ardu de classer précisément des textes sur une échelle de difficulté à plusieurs niveaux. Ils envisagent la possibilité d'utiliser à la place un classifieur binaire apte à décider si un texte est plus ou moins facile qu'un autre. Nous allons détailler chacune de ces trois directions dans la suite de cette section.

La lisibilité computationnelle lexico-syntaxique

Collins-Thompson et Callan (2004a, 2005) reprennent le modèle unigramme de Si et Callan (2001) auquel ils apportent certaines modifications. Tout d'abord, ils l'adaptent à tout type de textes et à l'ensemble du continuum primaire et secondaire, mais délaissent la variable syntaxique « longueur des phrases ». Leur principal apport réside dans une version lissée du modèle unigramme, intégré au sein d'un classifieur Naïf Bayésien multinomial.

Le principe est le suivant. À partir de 550 documents, issus de sites à vocation pédagogique et dont la difficulté a été évaluée par le concepteur de la page, ils entraînent 12 modèles de langue unigrammes. Autrement dit, chacun de ces modèles représente la distribution multinomiale d'un vocabulaire V au sein de la classe G_i, c'est-à-dire qu'à chaque mot w_j de V est associée une probabilité $P(w_j|G_i)$. Dès lors, classer un texte T revient à choisir la classe G_i qui maximise la quantité suivante :

$$L(G_i\,|T) = \sum_{w \in T} C(w) \log P(w\,|G_i) \tag{2.50}$$

où la probabilité $P(w\,|G_i)$ est lissée en fonction de la fréquence d'apparition du mot w dans les niveaux adjacents. C'est ce dernier aspect qui est particulièrement novateur par rapport à Si et Callan (2001) et qui, d'après les auteurs, améliore grandement l'exactitude. En validation croisée à 10 échantillons, ils obtiennent une corrélation avec leur critère de $0,79$ pour l'anglais et de $0,64$ pour le français. Cette chute de performance, d'autant plus étonnante que ce second modèle ne contient que cinq classes, s'explique possiblement par un manque d'adéquation du modèle au problème.

Parallèlement à ces travaux, une autre équipe américaine – de la *Washington University* – adoptent une approche similaire afin d'évaluer la difficulté de textes pour des lecteurs de l'enseignement traditionnel, mais dont l'anglais est la L2. Schwarm et Ostendorf (2005) préfèrent, au modèle unigramme lissé de Collins-Thompson et Callan (2004a), un modèle trigramme où la difficulté d'un texte est assimilée à sa perplexité (*cf.* Chapitre 3.2.2). Par ailleurs, le recours au parseur de Ratnaparkhi (1996) permet aux auteurs de disposer d'une description de la structure syntaxique et, par voie de conséquence, de définir des indices de la complexité syntaxique des phrases.

Ces diverses variables, ainsi que trois autres plus classiques, sont combinées à l'aide de SVM (*cf.* Chapitre 5) . Le corpus d'entraînement comprend environ 2400 articles d'un journal éducatif, le *Weekly Reader*, et leur difficulté se répartit de la deuxième à la cinquième primaire. Les résultats de ce modèle se comparent plutôt favorablement à ceux qu'obtiennent les formules de Kincaid *et al.* (1975) et de Stenner *et al.* (1988) sur un même corpus de test. Toutefois, on soulignera que le problème de classification ne compte que quatre classes, ce qui représente une tâche moins complexe que celle entreprise par Collins-Thompson et Callan (2004a). De plus, les auteurs emploient une nouvelle mesure d'évaluation, **l'exactitude contiguë**, qui correspond au pourcentage de textes mal classés de plus d'une classe par rapport au niveau de référence. Il va sans dire que, sur un problème à quatre classes, ce type de mesure se comporte plutôt optimistement.

Petersen et Ostendorf (2006a, 2009) poursuivent ces recherches en affinant le classifieur

et en approfondissant certaines problématiques. Ils remarquent que le classifieur de 2005 classe la majorité des textes qui lui sont fournis, même lorsque leur complexité dépasse largement le niveau de la cinquième primaire. Autrement dit, le modèle n'est pas généralisable à des textes d'une difficulté quelconque. Afin d'éviter ce désagrément, Petersen et Ostendorf (2009) ajoutent un ensemble d'articles destinés à des adultes au corpus d'entraînement. Ceux-ci se comportent comme une cinquième classe, correspondant à l'ensemble des textes plus ardus que ceux de la cinquième primaire, ce qui engendre une meilleure discrimination parmi les textes qui composent leur corpus de test. Dès lors, ils atteignent une exactitude de 43% pour cette tâche à 5 classes. Par ailleurs, leurs expériences semblent montrer que les variables syntaxiques participent peu aux performances globales du modèle.

C'est cette même problématique qui motive Heilman *et al.* (2007), lesquels postulent que si l'apport des prédicteurs syntaxiques reste incertain, il devrait être plus important lorsqu'il s'agit de prédire la lisibilité de textes en L2. Les auteurs reprennent le modèle unigramme lissé de Collins-Thompson et Callan (2004a) (ci-dessous L_{LM}) et le combinent linéairement avec le résultat d'un k plus proches voisins (kNN) (Hastie *et al.*, 2001) (où $k = 12$) exclusivement basé sur des variables syntaxiques (*cf.* aussi Section 2.3). Ce second modèle présente l'avantage de produire en sortie un intervalle de confiance C_{kNN} qui est associé à la classe L_{GR} attribuée au texte à analyser. Dès lors, l'influence du second modèle est pondérée au cas par cas selon la formule suivante (Heilman *et al.*, 2007, 464) :

$$\text{Prédiction} = L_{LM} + C_{kNN} * L_{GR} \tag{2.51}$$

Heilman *et al.* (2007) expérimentent ce modèle sur deux corpus : le premier rassemble 362 textes de L1 étiquetés par des professeurs en fonction des douze niveaux scolaires américains ; le second est constitué de 50 passages tirés d'une anthologie de textes à destination d'apprenants de l'anglais L2 et dont la difficulté s'échelonne de débutant (niveau 2) à avancé (niveau 5). Ils observent effectivement de meilleures performances pour le modèle sur la L2 que celui pour la L1, lequel a pourtant été entraîné sur davantage de textes.

Dans un article postérieur, Heilman *et al.* (2008) sont les premiers à envisager diverses échelles pour mesurer la difficulté (nominale, ordinale ou continue) et à comparer les performances d'algorithmes statistiques adaptés à ces trois types de variables dépendantes. Ils concluent à la supériorité de la régression logistique ordinale sur la régression linéaire et la régression logistique multinomiale. Par la suite, Kate *et al.* (2010) parviennent pourtant à un résultat radicalement opposé : leurs expériences favorisent les algorithmes à base de régression et, en particulier, le bagging à base d'arbres de décision (*cf.* Chapitre 5).

Par ailleurs, leur étude comporte quelques autres aspects intéressants. Tout d'abord, ce sont

les premiers à constituer un corpus qui mélange des textes rédigés par des humains avec d'autres générés par une machine. Ces documents ont été annotés par huit experts selon une échelle de 1 à 5. Si leurs variables ne sont pas révolutionnaires – elles restent lexico-syntaxiques et s'inspirent des études décrites précédemment –, Kate *et al.* (2010) sont également les premiers à effectuer une expérience qui tend à prouver que les prédictions d'un ordinateur surpassent celles d'humains non spécialisés. Cette observation est évidemment essentielle, car elle justifie non seulement l'approche automatisée de la lisibilité, mais aussi son usage pour définir le niveau de textes destinés à l'instruction.

La prise en compte des aspects structuro-cognitifs

À ce jour, les synthèses des paradigmes classiques et structuro-cognitifs restent rares, même si la tendance risque de s'inverser à mesure que la linguistique computationnelle effectuera des avancées dans le domaine de la sémantique. On peut considérer Graesser *et al.* (2004) comme des précurseurs en la matière. En effet, leur logiciel Coh-Metrix est capable de paramétriser automatiquement un grand nombre de dimensions textuelles, aussi bien lexicales et grammaticales que discursives. Parmi celles-ci, les plus intéressantes sont l'âge d'acquisition et l'imageabilité des mots, la polysémie, et, en particulier diverses mesures de la cohésion et de la co-référence [59]. Cette étude a le mérite de porter à l'attention des chercheurs en lisibilité un grand nombre de facteurs issus de la psychologie cognitive qui restent encore trop souvent négligés. La preuve en est que, n'ayant pris connaissance de ces travaux que trop tardivement, nous avons redécouvert certains de ces facteurs au cours de nos propres recherches.

Il faut attendre trois ans avant que Crossley *et al.* (2007) ne combinent ces différentes variables détectées par Coh-Metrix et proposent la première formule TAL complètement automatisée qui envisage à la fois facteurs traditionnels et discursifs. En réalité, la démarche reste assez traditionnelle : les auteurs entraînent leur formule sur les 32 passages de Bormuth, ce qui représente un nombre réduit d'observations. Dès lors, afin d'éviter un surentraînement, ils se limitent à trois facteurs : un lexical (la fréquence des mots), un syntaxique (le nombre de mots par phrase) et un cohésif (l'*argument overlap*, c'est-à-dire le nombre de paires de phrases comportant chacune un nom dont la racine est la même). Combinées par régression linéaire, ces trois variables expliquent 91% de la variance, ce qui rejoint les performances de la formule de Chall et Dale (1995), également développée sur le même corpus. Toutefois, l'ajout du facteur cohésif, que les auteurs présentent comme une avancée majeure et un moyen de répondre aux critiques des années 80, ne convainc pas complètement. D'une

59. La liste complète et détaillée de ces caractéristiques est disponible à l'adresse `http://cohmetrix.memphis.edu/CohMetrixWeb2/HelpFile2.htm`.

part parce que la formule de Chall et Dale (1995) s'en passe très bien et obtient des résultats similaires ; d'autre part parce que ce facteur, une fois combiné avec les deux autres, ne se révèle plus vraiment significatif ($p = 0,062$).

L'année suivante, Pitler et Nenkova (2008) poussent cette démarche plus loin. Elles disent elles-mêmes vouloir présenter un modèle de lisibilité unifié. Pour ce faire, elles sélectionnent 30 articles du Wall Street Journal dans le Penn Discourse Treebank (Prasad *et al.*, 2008), lequel propose des textes où le type de relations entre chaque phrase et leur sens ont été annotés. La difficulté de chacun de ces 30 textes est évaluée par trois étudiants sur une échelle de 1 à 5. Au niveau des prédicteurs, les auteurs en ont implémenté trente-deux, parmi lesquels on retrouve quatre variantes du modèle unigramme de Collins-Thompson et Callan (2004a), les quatre facteurs syntaxiques de Schwarm et Ostendorf (2005) et une série de mesures discursives et cohésives. Celles qui sont le plus corrélées avec les annotations des étudiants sont la log-vraisemblance d'un article en fonction de ses relations discursives ($r = 0,4835$), le modèle unigramme ($r = 0,4497$) et le nombre moyen de syntagmes verbaux par phrase ($r = 0,4213$). Autrement dit, on retrouve un facteur pour chacune des trois principales dimensions textuelles et, cette fois, l'importance des facteurs structuro-cognitifs en lisibilité est confirmée.

En combinant huit variables, les deux chercheuses atteignent un R de $0,88$ et remarquent que parmi les facteurs retenus, certains ne sont pourtant pas corrélés significativement avec le critère. Elles mettent ainsi en évidence la difficulté de combiner des variables indépendantes. On peut d'ailleurs craindre un surentraînement du modèle étant donné le peu de documents utilisés pour effectuer cette opération et l'absence de résultats en validation croisée. Soulignons aussi l'aspect peu pratique de l'échelle employée : elle ne correspond pas à une norme répandue et risque de rebuter les utilisateurs potentiels.

Dans la lignée de ces deux études, Feng *et al.* (2009) s'interrogent sur les moyens de prédire la compréhension de documents par des adultes atteints de déficiences intellectuelles. Ils expérimentent 25 variables, parmi lesquelles on compte trois variables traditionnelles, des variantes des facteurs syntaxiques décrits par Petersen et Ostendorf (2009) et, surtout, une série de variables motivées par des justifications cognitives. En effet, les personnes atteintes de déficiences intellectuelles rencontreraient des problèmes de lecture à cause de limitations mémorielles et d'un encodage sémantique ralenti. Dès lors, Feng *et al.* (2009) définissent un ensemble de variables basées sur les « entités » (qui rassemblent les noms propres et noms communs) et sur la notion de chaînes lexicales, décrites par Galley et McKeown (2003). Ces facteurs visent à mieux modéliser la charge mémorielle d'un texte.

À partir de ces trois familles de caractéristiques, les auteurs entraînent trois formules sur des articles du *Weekly Reader*, déjà utilisés par Petersen et Ostendorf (2009). La première d'entre

elles ne recourt qu'à des variables bien connues dans la littérature ; la seconde se limite aux nouveaux facteurs cognitifs implémentés par les auteurs, tandis que la troisième mêle ces deux types d'informations. On observe un résultat étonnant. Sur un échantillon de test également issu du *Weekly Reader*, la formule qui combine des variables de surfaces et des variables cognitives est la plus précise. Par contre, c'est la formule purement cognitive qui se comporte le mieux vis-à-vis de vingt autres textes, dont la difficulté a été estimée au moyen de tests de compréhension soumis à des personnes atteintes de déficiences intellectuelles. Les auteurs expliquent ce résultat par le fait que les problèmes éprouvés par ce type de public ne se situent pas au niveau du décodage, mais plutôt de niveaux de traitement supérieurs, plus adéquatement mesurés par les variables cognitives.

Abandonnant toutefois ce public particulier dans un article postérieur, Feng *et al.* (2010) y proposent une synthèse des variables employées en lisibilité par la linguistique computationnelle. Ils réutilisent le corpus du Weekly Reader, retenant 1433 articles répartis selon quatre niveaux de difficulté, pour effectuer une série d'expériences visant à identifier les meilleures variables (parmi 273) ainsi que le type d'algorithme statistique à privilégier. Ainsi, ils notent que (1) les algorithmes de classification se comportent nettement mieux que ceux basés sur la régression linaire et (2) que les classifieurs ordinaux, défendus par Heilman *et al.* (2008), obtiennent des performances identifiques ou pires que celles d'algorithmes standards. Par contre, les aspects structuro-cognitifs ne semblent guère apporter d'informations utiles pour définir une formule de lisibilité, du moins lorsque les textes considérés sont trop simples. C'est le cas ici, puisque les textes du corpus sont destinés aux primaires. Les variables dotées du plus grand pouvoir prédictif sont les catégories du discours, en particulier, les noms ; la proportion de syntagmes verbaux, un résultat consistant avec ceux de Pitler et Nenkova (2008) ; la longueur moyenne des phrases et le modèle de langue.

Approche par classification binaire

Parmi les autres directions explorées actuellement en lisibilité, on trouve quelques articles qui posent le problème sous une autre forme. Partant du principe qu'il est ardu de mettre au point un algorithme discriminant correctement plusieurs niveaux, ils préfèrent se limiter à une tâche binaire. Autrement dit, il s'agit de déterminer si un texte est plus difficile qu'un autre. Pitler et Nenkova (2008, 193) explorent les premiers cette voie. À partir des 30 articles de leur corpus, ils retiennent toutes les paires de textes a et b dont l'écart de niveau est supérieur à un seuil $\theta = 0,5$ et fusionnent les deux vecteurs de caractéristiques de ces deux textes sous la forme d'un vecteur unique (V_{ab})[60]. Ensuite, sur la base de ces vecteurs

60. Tanaka-Ishii *et al.* (2010, 208) décrivent trois techniques de fusion possibles : la simple concaténation, la conjonction via une soustraction ou via une division. Ils postulent que Pitler et Nenkova (2008) ont employé la soustraction, bien que ce ne soit pas précisé par les auteurs.

uniques, ils entraînent, à l'aide d'un SVM, une fonction de comparaison capable de détecter si a est plus difficile ou plus facile que b. Ce faisant, Pitler et Nenkova (2008, 193) atteignent une exactitude de $88,88\%$.

Un premier avantage de cette démarche est qu'elle permet de disposer, à partir d'un corpus de n documents, de $2 * n * n$ documents, et partant de réduire nettement les problèmes d'entraînement. Plus remarquable encore, Tanaka-Ishii *et al.* (2010) montrent que considérer ce type de modèle discriminant comme une fonction de tri permet d'ordonner efficacement un ensemble de textes en les comparant deux à deux. Cette démarche ne convient évidemment que pour ordonner des documents de manière relative et non pour prédire leur niveau dans l'absolu. Les auteurs obtiennent d'excellents résultats pour le japonais (L1 et L2) et l'anglais (L1 et L2), aussi bien pour la simple tâche de discrimination que pour le tri, dont les résultats sont corrélés avec une annotation de référence. En particulier, leurs expériences révèlent que cette approche par tri fonctionnerait même mieux qu'une classification par SVM. Toutefois, les auteurs restent modestes. Ils doutent que cette supériorité se maintienne lorsqu'on envisage davantage de facteurs et qu'on dispose d'un corpus plus large pour entraîner le classifieur.

Notons enfin, dans une veine tout à fait traditionnelle, les efforts de Weir et Ritchie (2006). Ils combinent linéairement deux facteurs classiques – le logarithme des fréquences (basées sur le British National Corpus (BNC)) et la longueur des phrases – au sein d'une formule implémentée en Java. Les auteurs présentent la prise en compte des fréquences lexicales comme une avancée, qui permettrait de mieux modéliser la familiarité avec la langue. Rappelons que ce facteur était déjà utilisé par Stenner (1996) et qu'il constitue aujourd'hui plutôt une baseline qu'une avancée. Par ailleurs, les auteurs reprennent l'hypothèse de Damerst (1972) selon lequel il existe une longueur optimale pour les phrases et la raffinent. Ils considèrent en effet que des phrases courtes doivent être considérées comme un obstacle à la compréhension lorsque le vocabulaire du texte est complexe, mais pas s'il est simple. La même année, Anagnostou et Weir (2006) proposent une idée plus novatrice : prendre en compte la fréquence moyenne des collocations détectées dans le texte. Malheureusement, ils ne rapportent pas de formules, ni de données empiriques validant ou infirmant cette hypothèse.

Conclusion

Lorsqu'on analyse cette dernière période, on est frappé par les évolutions qui ont eu lieu en moins de dix ans. Tout d'abord, la méthodologie traditionnellement employée depuis Vogel et Washburne (1928) a volé en éclat : évaluer la difficulté d'un texte n'est plus considéré uniquement comme un problème de régression, mais aussi comme une tâche de classifica-

tion, ce qui offre une grande variété d'algorithmes statistiques au chercheur. Parmi ceux-ci, aucun ne semble encore faire l'unanimité, mais plusieurs travaux ayant comparé divers algorithmes (François, 2009b; Kate *et al.*, 2010) laissent entendre que la plupart d'entre eux sont relativement équivalents en termes de performance et que le choix des indices de la difficulté serait davantage déterminant. À ce niveau, les outils de TAL existant ont facilité la détection automatisée de variables qui n'étaient jadis manipulables que via un traitement manuel. Ainsi, les formules disponibles aujourd'hui incluent un grand nombre de ces facteurs structuro-cognitifs dont l'absence était si critiquée dans les années 80.

Il apparaît donc qu'il y ait tout lieu d'être optimiste quant à l'avenir de la lisibilité. Le nombre des publications ne cesse de croître, de même que le nombre de domaines d'application directement concernés par le développement de formules automatisées. On peut ainsi les utiliser pour déterminer le naturel ou la complexité de textes générés par un ordinateur (Antoniadis et Grusson, 1996), produits par un traducteur automatique ou résumés automatiquement (Aluisio *et al.*, 2010; Kanungo et Orr, 2009; Wan *et al.*, 2010), mais aussi pour ordonner les résultats d'un moteur de recherche.

Toutefois, il nous semble opportun de modérer cet enthousiasme par plusieurs observations. La première d'entre elles concerne le glissement disciplinaire qu'a connu la lisibilité. Autrefois une matière à la croisée de la psychologie de l'éducation et de la communication, elle dépend désormais plutôt des sciences cognitives, voire des sciences informatiques. Nous voyons dans ce glissement la possibilité d'une perte de contact avec les problématiques de terrain ou d'une diminution de l'expertise sur les questions éducatives propres à la lecture. Ainsi, de moins en moins de chercheurs utilisent des échelles directement opératoires pour les professeurs, telles que les *grades* américains. Ils leur préfèrent des annotations plus simples, comme une échelle allant de 1 à 5. Dans le même ordre d'idée, l'interrogation qui a hanté les chercheurs des générations précédentes, à savoir si le critère retenu mesure adéquatement la compréhension textuelle, semble actuellement occultée. La grande majorité des chercheurs recourent à des corpus annotés par des experts. Ce qui compte désormais est d'obtenir la meilleure performance par rapport à un golden standard, au risque de se perdre dans des optimisations compliquées qui se révèlent finalement peu significatives.

Comme au temps de la révolution cognitiviste, on observe chez certains un oubli relatif de la tradition en lisibilité. Le cas le plus flagrant est Weir et Ritchie (2006) qui omettent dix ans de travaux en présentant le calcul des fréquences comme une approche innovante. Mais des chercheurs aussi reconnus que Schwarm et Ostendorf (2005) dénient l'efficacité des variables traditionnelles telles que la longueur moyenne des phrases et mettent en avant l'emploi d'un parseur statistique. Nulle part dans leur article, on ne trouve de référence aux expériences de Bormuth (1966) qui a longuement étudié les variables basées sur l'analyse

syntaxique et qui en a conclu qu'elles ne constituent pas de meilleurs prédicteurs que la longueur des phrases. Plus généralement, nous avons remarqué que ce sont souvent les mêmes références qui sont présentées dans les différents états de l'art des publications actuelles. Il s'agit essentiellement de Flesch (1948), Dale et Chall (1948a), Kincaid *et al.* (1975) ou encore Gunning (1952).

Pour le français, la situation est bien plus sombre. En effet, il n'existe, à notre connaissance, aucune recherche publiée sur la lisibilité du français L1 depuis celle de Mesnager (2002). Il faut dire que la quantité d'applications TAL disponibles pour cette langue est nettement plus réduite que pour l'anglais, sur lequel portent la majorité des recherches actuelles. Ainsi, il n'est pas évident que les diverses variables basées sur une analyse syntaxique soient applicables au cas du français, pour lequel il manque toujours un analyseur syntaxique fiable et disponible à des fins de recherche. De même, la plupart des facteurs discursifs reposent sur des formalismes et des outils développés sur l'anglais et qui demandent une adaptation. La tentation de se limiter à cette langue est donc grande, d'autant que si quelques corpus annotés existent en anglais et sont référencés dans la littérature, ce n'est pas le cas du français.

Pour conclure, nous pensons qu'il est plus que probable que l'enthousiasme actuel perdure encore plusieurs années. Parmi les différentes voies envisagées dans le monde anglo-saxon, la plus prometteuse nous apparaît être l'approche mixte, qui combine des facteurs de tous types, de surface comme des discursifs. Lancée par Foltz *et al.* (1998), cette démarche a été poursuivie par Crossley *et al.* (2007), Pitler et Nenkova (2008) et Feng *et al.* (2009, 2010). Espérons qu'elle débouche sur des applications également pour le français.

2.2.9 Conclusion sur les études en L1 : un parcours sinueux

Au cours de ce long état de l'art des recherches en lisibilité, portant majoritairement sur l'anglais L1 et, dans une moindre mesure, sur le français L1, nous avons vu se répéter une suite de cycles. Chacun de ces cycles commence par une innovation majeure, parfois issue du rapprochement avec une autre discipline, qui stimule la discipline et déclenche une période faste où de nouvelles pistes méthodologiques sont mises au jour. Peu à peu, celles-ci sont explorées et montrent, au bout d'un moment, leurs limites. Dès lors, les doutes et les critiques, occultées un moment, reviennent au goût du jour et la créativité des auteurs diminue. Ceux-ci se limitent souvent, soit à faire le procès de la lisibilité sans apporter de solutions, soit à optimiser les techniques récentes sans remettre en question les paradigmes utilisés.

Ainsi, malgré la publication d'études séminales (Gray et Leary, 1935; Ojemann, 1934; Vogel et Washburne, 1928) qui mettent en place une méthodologie qui survit jusqu'à l'aube du 21^{e}

siècle, la lisibilité reste un domaine confidentiel à ses débuts. Ce n'est que lorsque les leçons de McCall and Crabbs sont utilisées par Lorge (1939) et reprises par Flesch (1948) et Dale et Chall (1948a) que le domaine perce dans la sphère publique. Toutefois, les recherches tournent assez vite en rond, abordant de multiples façons les deux mêmes types de facteurs. Aussi, la période qui va de 1953 à 1965 n'offre guère de nouveautés, si ce n'est les premières confrontations avec un ordinateur (Smith, 1961).

Il faut attendre Coleman et l'arrivée du test de closure pour que s'initie un nouveau cycle au cours duquel Bormuth (1966, 1969a) affirme les bases théoriques du domaine, tandis que le recours à l'ordinateur s'intensifie. Toutefois, d'autres (Carver, 1974a; Singer, 1975) envisagent des chemins de traverse, où ils espèrent éviter les ornières méthodologiques qui se présentent toujours de la même manière aux chercheurs. Ces ornières sont d'ailleurs à l'origine d'une nouvelle vague de critiques, de la part des structuro-cognitivistes, qui se révèlent plus acerbes que leurs prédécesseurs. Ce paradigme ne se contente toutefois pas de critiquer : il apporte avec lui nombre de connaissances en grammaire des textes et en psychologie cognitive de la lecture qui redynamisent le domaine, initiant un troisième cycle.

Malheureusemenent, il apparaît assez vite que ces connaissances, bien que cruciales pour expliquer les origines des difficultés de compréhension à la lecture, sont encore inappropriées pour prédire cette difficulté à l'aide de moyens pratiques. Le domaine entre alors dans une phase de dormance entre la fin des années 80 et le début du 21^{e} siècle, durant laquelle peu de travaux paraissent. On a presque l'impression que la lisibilité est oubliée, avant que le mariage avec la linguistique computationnelle ne relance l'intérêt de la communauté scientifique.

Nous assistons actuellement au début d'un quatrième cycle et à une revivification du domaine. Du moins en ce qui concerne l'anglais, le nombre de travaux publiés chaque année ne cesse de croître. Toutefois, il n'est pas impossible que la lisibilité coure le risque que nous avons décrit ci-dessus : celui d'être coupé de ses racines. C'est pourquoi, dans le cadre de cette thèse, nous avons voulu dresser le portrait le plus complet possible de ce long parcours historique et scientifique à la poursuite d'une sorte de pierre philosophale : un modèle capable de prédire la difficulté des textes. Il s'agit, à notre connaissance, de la synthèse la plus complète dans le domaine, et qui est également la seule disponible en français. Nous espérons que le lecteur ne nous tiendra pas rigueur de sa longueur et pour conclure ce parcours, nous terminons ce chapitre en envisageant la question spécifique de la lisibilité en langues étrangères, en particulier en ce qui concerne le FLE.

2.3 La lisibilité pour les langues étrangères

2.3.1 Introduction à la lisibilité en langues étrangères

La lisibilité en langues étrangères partage le même objectif que la lisibilité en L1, c'est-à-dire évaluer le niveau de difficulté de textes à l'aide de formules mathématiques. Pendant longtemps, les deux domaines n'ont pas vraiment été distingués l'un de l'autre. Il était coutume d'employer une formule conçue pour des natifs pour évaluer le niveau des textes destinés à des apprenants d'une L2. Ce faisant, on suivait le raisonnement décrit par Greenfield (2004, 8) :

> *Readability formulas assume that the readers for whom they predict difficulty and the texts to which they are applied are similar to the samples used to derive the formulas in the first place. [...]. Using the formulas to predict difficulty for second language readers assumes that those readers are not significantly different from native readers in ways that affect how the measured text variables relate to reading difficulty.*

Or, dans le Chapitre 1, nous avons montré qu'il existait des différences notables entre les processus de lecture en L1 et en L2. Les apprenants d'une L2 sont plus sensibles au lexique et à la syntaxe, du moins tant qu'ils n'ont pas atteint le fameux seuil au-delà duquel ils sont supposés pouvoir transférer leurs aptitutes à la lecture de leur L1 vers la L2. Il est donc douteux que les formules développées pour des natifs soient valides pour des apprenants d'une L2. Cette question, pourtant cruciale, n'a pourtant fait l'objet que d'un nombre restreint d'études qui, comme nous allons le présenter, s'accordent difficilement entre elles.

De manière plus générale, la lisibilité des L2 reste un domaine peu couru où les publications sont rares. Il ne faut d'ailleurs pas confondre les études sur la L2, qui comprennent l'anglais, avec les travaux en L1 portant sur d'autres langues que l'anglais. Le risque de confusion est d'autant plus grand que les principales revues sur la question (Klare, 1974; Rabin, 1988) traitent de ces deux aspects.

Dans la suite de notre propos, nous décrivons tout d'abord les premiers travaux en lisibilité des langues étrangères. Touchant aussi bien au français, à l'anglais qu'à l'espagnol, ils se caractérisent par un intérêt quasi exclusif pour les aspects lexicaux. Dans un second temps, nous couvrons le peu de formules classiques dont nous avons eu connaissance pour ces trois langues, avant de distinguer, au sein des recherches récentes, ce qui concerne l'anglais comme langue seconde (ALS) et le FLE. On notera, qu'une fois de plus, la plupart des travaux présentés et, en particulier les plus anciens d'entre eux, ont été développés aux États-Unis.

2.3.2 Les premiers travaux

En ce qui concerne la lisibilité des matériaux pour l'enseignement des langues étrangères, Venezky (1984, 24) rapporte que, comme pour les études de lisibilité en L1, c'est au niveau du vocabulaire qu'ont lieu les premières tentatives de contrôle de la difficulté. Dès les années 1910, des listes de mots auxquels étaient associés des comptages sont établies pour le latin, le russe ou l'allemand. Leur existence a probablement incité Thorndike à compiler une liste similaire pour l'anglais L1, son *Teacher's Word Book*.

Suite à l'apparition de ce type de listes, plusieurs chercheurs les ont employées pour tenter de mesurer plus objectivement la difficulté de textes en langue étrangère. Ainsi, Johnson (1927) réduit la liste de fréquence de Henmon, *A French Word Book*, de 3900 à 1695 termes à l'aide d'une formule qui combine la fréquence d'occurrence et le nombre de sources d'un mot, suite à quoi il calcule, pour un ensemble de textes, le pourcentage de mots absents de sa liste. Il observe des différences importantes entre les différents livres de son échantillon et conclut à l'efficacité de la « fréquence » des mots comme prédicteur des difficultés de lecture. Toutefois, ces difficultés ne sont pas mesurées à l'aide d'un critère objectif. C'est ici le bon sens des professeurs qui fait office de critère.

Dans le même ordre d'idée, Arnold (1932) analyse le lexique utilisé dans 15 œuvres espagnoles sur la base de la liste de Buchanan (*Spanish Word Book*). Il démontre qu'on obtient des estimations plus fiables en travaillant avec les *types* plutôt que directement avec les occurrences. Il dresse pour chaque livre un profil lexical, selon une démarche qui rappelle Dolch (1928). Ce profil détaille le pourcentage de mots présents dans le premier millier de mots de la liste de Buchanan, dans le second millier, etc. ainsi que le pourcentage d'absents. Par ailleurs, Arnold (1932) est le premier – à notre connaissance – à faire allusion au concept de congénères, ces mots d'une L2 dont le sens est similaire et la forme homographique ou presque avec leurs équivalents dans la L1 du lecteur. Leur présence peut faciliter la lecture d'un texte en langue étrangère et mettent à mal l'association entre fréquence et difficulté des mots.

Cette technique basée sur le pourcentage de mots appartenant à chaque tranche de mille mots est aussi utilisée, pour le français, par Cole (1932, cité par Tharp, 1939, 171) qui se base sur le *French Word Book* de Vander Beke. Stone et Mullins (1932) proposent une exploration plus systématique de différentes variables lexicales, en se basant aussi sur la liste de Vander Beke. Elles cherchent à combiner ces différentes variables en une mesure unique capable d'attribuer un indice de difficulté à un texte, et font même intervenir des aspects syntaxiques tels que le pourcentage de phrases simples, complexes ou composées. Si l'on distingue de nombreux points communs avec la tradition en lisibilité de l'anglais, en particulier en ce qui

concerne les indices de la difficulté, il apparaît clairement que ces travaux manquent d'un cadre méthodologique similaire à celui utilisé par Vogel et Washburne (1928) ou Ojemann (1934) à la même époque. Par contre, on rencontre déjà certaines observations propres au contexte L2.

Tharp (1939) va synthétiser ces premiers résultats, qui découlent pour lui des travaux de Lively et Pressey (1923). Il leur reproche les faiblesses suivantes :

1. Tout d'abord, au sein d'une tranche de x mots, la difficulté est considérée comme constante, alors que la charge lexicale suit plus probablement un continuum.
2. Si la démarche de Lively et Pressey (1923) semble valide pour des lecteurs en L1, elle pose question en L2 à cause des congénères. Ainsi, il n'est pas normal que le dernier mot sur la liste de Vander Beke soit « zéro », alors que ce terme est un congénère parfait pour un anglophone qui le comprendra aisément. Cette observation est valable pour l'ensemble des congénères, pour autant que leurs équivalents soient connus par un apprenant donné dans sa langue maternelle.
3. De plus, au sein des différentes flexions d'une même forme canonique, il est nécessaire de distinguer, celles qui sont transparentes pour les apprenants de celles qui posent des problèmes.
4. De la même façon, il conviendrait de distinguer les différents sens d'une même forme graphique, étant donné que leur complexité peut varier considérablement.
5. Enfin, Tharp rappelle qu'il importe aussi de considérer la densité des mots représentant un fardeau pour la compréhension (*burden words*), lesquels varient évidemment d'un individu à l'autre. En effet, il est plus simple de traiter un mot inconnu lorsqu'il est environné de nombreux termes connus que lorsqu'il côtoie d'autres vocables tout aussi abscons.

Dès lors, Tharp (1939) et ses étudiants réalisent les deux expériences suivantes. D'une part, ils sélectionnent, pour huit textes français, les *burden words* (B) dont la densité est calculée comme le nombre de mots du texte divisé par le nombre de *burden words*. De ces termes complexes sont exclus les noms propres et les congénères dont l'équivalent anglophone est présent dans les 10 000 premiers mots de la liste de Thorndike. En sus de la densité, les auteurs segmentent les 6 136 mots de la liste de Vander Beke en 12 catégories et dénombrent le pourcentage de mots tombant dans chacune de ces tranches. Afin d'obtenir un indice unique de la charge lexicale, ils établissent la formule suivante :

$$\text{Frequency Index} = \frac{\sum_{i=1}^{13} F * i}{T} \tag{2.52}$$

où T représente le nombre total de mots du texte à analyser, F correspond à la fréquence absolue des mots qui appartiennent à la tranche C_i, tandis que i est la pondération appliquée à

F. Les mots appartenant à des classes moins fréquentes pèsent ainsi davantage dans l'indice de fréquence. Notons que la section C_{13} correspond aux mots absents de la liste. On peut évidemment discuter de la pertinence du découpage selon ces 12 tranches et du système de pondération que l'auteur justifie par des considérations pratiques. De même, leur tentative de combiner la densité et l'indice de fréquence en un ratio ne semble pas conduire à des résultats exceptionnels. Pour autant, Tharp a défini des critères que devrait respecter toute formule de lisibilité destinée à des textes en L2.

2.3.3 La période classique

Après cette jonction entre les traditions en L1 et L2 opérée notamment par Tharp (1939), les publications postérieures adoptent la méthodologie classique. Malheureusement, ce rapprochement conduit apparemment le milieu éducatif à estimer que les formules L1 d'une langue sont aussi valides pour un contexte L2, ce qui explique probablement la rareté des études spécifiques à la L2. Spaulding (1951) constitue une exception à cette vision. Bien qu'il emploie l'approche classique pour développer ses deux formules pour l'espagnol L2, il ne néglige pas les particularités du contexte L2.

Il fait évaluer la complexité de 20 textes par des experts et analyse la corrélation entre cette échelle et quatre variables : le nombre moyen de lettres par mot ; le nombre moyen de mots par phrase (ASL) ; un indice de fréquence, similaire à celui de Tharp (1939)[61] (FI) et la densité lexicale (D), définie comme le nombre total de mots divisés par le nombre de mots ardus, c'est-à-dire n'étant ni un congénère, ni l'une des 1 500 premières entrées de la liste de Buchanan. Il observe que les deux variables les plus corrélées avec le critère sont ASL $(0,788)$ et FI $(0,688)$ et les combine en une formule (F_1). Toutefois, comme le calcul de FI est fastidieux, il propose une seconde formule F_2 où celui-ci est remplacé par la densité (D). Les deux formules sont les suivantes (Spaulding, 1951, 124) :

$$\begin{aligned} F_1 &= -2,383 + 4,115\, FI + 0,154\, ASL \\ F_2 &= 2,20 + 33,18\, D + 0,1609\, ASL \end{aligned} \tag{2.53}$$

La différence de performance entre F_1 ($R = 0,91$) et F_2 ($R = 0,87$) est suffisamment faible pour pousser Spaulding à recommander uniquement la seconde dans un article postérieur (Spaulding, 1956) où se trouvent également les instructions précises pour appliquer F_2 à un texte. Un dernier élément intéressant à relever est la différence de corrélation qu'entre-

61. Plus exactement, cet indice est calculé de la façon suivante. Après que la liste de Buchanan a été segmentée en 13 classes C_i d'environ 500 mots, l'indice de fréquence est calculé selon l'équation 2.52 où $1 \leq i \leq 14$. Il faut noter que les congénères rencontrés se voient attribuer un poids de 1, quelle que soit leur position dans la liste de Buchanan.

tient l'indice de fréquence avec le critère, suivant que l'on considère les congénères comme connus (c'est le cas par défaut) ou d'un niveau de difficulté équivalent à leur fréquence dans la L1 (tranche i dans Buchanan). Dans le premier cas, la corrélation atteint $0,688$ pour $0,62$ dans le second. Ce résultat semble valider empiriquement l'hypothèse théorique de Tharp (1939) qui prend en compte de l'effet de facilité dû aux congénères.

La diffusion de ce constat semble toutefois être resté confidentielle, puisque Laroche (1979) fait paraître un article dans lequel il critique l'application d'une formule de L1 à des textes en L2, en décrivant les limites méthodologiques de cette façon de procéder. Ces remarques concernent notamment l'emploi de listes de fréquences. En effet, si l'on peut postuler que pour un natif, constamment immergé dans un bain linguistique, la fréquence d'un mot est effectivement lié à sa difficulté, rien ne permet d'affirmer que cette association reste valide pour un apprenant. Cette remarque correspond à une question cruciale sur laquelle nous reviendrons dans le Chapitre 3. Par ailleurs, il insiste sur la notion de congénères, dont il propose deux techniques de paramétrisation : un simple dénombrement ou une mesure de fréquence qui nécessitent toutes deux de développer une ressource spécifique à une paire de langues donnée.

Dans les différents travaux présentés jusqu'ici, on remarque que la difficulté de textes en L2 n'est souvent envisagée qu'au travers de la charge lexicale. Gaies (1979) établit le même constat à propos des principes régissant la progression des livres simplifiés disponibles sur le marché. Il déplore d'une part l'absence de standards définis en référence à des niveaux de compétence, ce qui conduit chaque éditeur à recourir à sa propre liste de vocabulaire. D'autre part, il estime que la syntaxe est négligée et propose comme mesure de la complexité syntaxique le nombre de mots par T-unit [62]. Quantifiant cette variable dans une vingtaine de livres simplifiés de différents niveaux, Gaies observe une sorte de corrélation (il ne rapporte pas de valeur) entre la complexité syntaxique et le niveau des textes, mais il ne poursuit pas la démarche plus avant.

2.3.4 Les recherches récentes pour l'anglais L2

La période classique, malgré quelques réflexions très pertinentes, ne produit guère d'avancées dans les formules particulières à la L2. On continue d'utiliser des outils développés pour la L1 ou, au mieux, de s'intéresser à la question des congénères. Il faut attendre le début des années 80 pour que soit à nouveau remise en question la validité de ce transfert des formules de L1. Ainsi, Hamsik (1984, cité par Greenfield, 2004) investigue la validité des formules de Flesch, Dale-Chall, Fry et Lorge à partir de 18 tests de closure soumis à

62. Cette unité syntaxique particulière a été définie par Hunt (1965). Elle correspond à une proposition principale accompagnée de l'ensemble de ses propositions subordonnées.

40 apprenants d'ALS d'origine diverses. Elle obtient des corrélations entre $0,775$ et $0,819$ pour ces quatres formules et conclut à leur validité dans le contexte de l'ALS, mais suggère la possibilité de concevoir des formules spécifiques à la L2 qui seraient plus précises. Étant donnée la petite taille de son échantillon, elle n'a pas pu étudier l'existence d'effets liés au background des apprenants.

Brown (1998, cité par Crossley *et al.*, 2008) réitère l'expérience de Hamsik à plus grande échelle. Il administre, à près de 2300 étudiants universitaires japonais, 50 textes sous la forme de tests de closure et compare les scores moyens de chaque texte avec six formules classiques dont celles de Flesch, Flesch-Kincaid, Gunning ou Fry. Les taux de corrélation qu'il obtient, compris entre $0,48$ et $0,55$, sont plutôt faibles, ce qui le pousse à rejetter la validité de formules pour natifs employées en L2. Il propose à la place une formule spécifique qu'il appelle *EFL* [63] *Difficulty Estimate* (Greenfield, 2004, 9) :

$$\text{EFL Difficulty} = 38,7469 + 0,7823\,X_1 - 126,1770\,X_2 + 1,2878\,X_3 + 0,7596\,X_4 \quad (2.54)$$

EFL Difficulty : le score moyen qu'obtiendrait le texte analysé à un test de closure administré à des apprenants d'ALS japonais ;

X_1 **:** le nombre de syllabes par phrase ;

X_2 **:** le nombre de fois que le terme supprimé apparaît ailleurs dans le même texte ;

X_3 **:** le pourcentage de mots de plus de six lettres ;

X_4 **:** le pourcentage de mots grammaticaux.

Cette équation obtient un R de $0,74$, une corrélation qui n'est pas particulièrement forte, mais qui surpasse les performances des six formules classiques testées sur le même corpus. Brown en conclut qu'il est préférable de développer des formules particulières pour chaque population d'apprenants de L2 dont le background diffère.

Dans le cadre de sa thèse, Greenfield (1999, cité par Greenfield, 2004) a effectué une étude similaire. Il compare les cinq formules suivantes – Flesch, Flesch-Kincaid, Coleman-Liau, Bormuth (1969a) et Chall et Dale (1995) –, avec les résultats de 200 étudiants japonais d'ALS à des tests de closure portant sur 30 des 32 passages du corpus de Bormuth. Parmi ces cinq compétiteurs, c'est la formule de Bormuth qui ressort, avec un R de $0,861$. Ce premier résultat soutient l'optique qui consiste à appliquer une formule de L1 à un contexte L2. Il est renforcé par une seconde observation. Greenfield a en effet corrélé les résultats obtenus par ses étudiants avec ceux qu'avaient obtenus les sujets natifs de Bormuth sur ces mêmes textes. La corrélation, très forte ($r = 0,915$), permet à l'auteur de conclure que les difficultés rencontrées par des natifs et des apprenants d'ALS sur ces textes sont relativement

63. Le sigle EFL est l'équivalent en anglais de ALS, que nous privilégions dans cette thèse.

similaires. Notons que ce constat n'est valide que si le test de closure mesure effectivement la compréhension des apprenants d'une langue étrangère.

Ne s'arrêtant pas là, Greenfield réentraîne les cinq formules susmentionnées sur son corpus, dont la difficulté correspond mieux à la réalité d'apprenants de L2. Le gain en termes de performances est toutefois minime. Dès lors, l'auteur propose sa propre équation, très simple, (Greenfield, 2004, 12) :

$$\text{Miyazaki EFL Readability Index} = 164,935 - 18,792\,X_1 - 1,916\,X_2 \tag{2.55}$$

Miyazaki EFL Readability Index : est le score moyen (compris entre 0 et 100) qu'obtiendrait le texte analysé à un test de closure administré à des apprenants d'ALS japonais ;

X_1 **:** le nombre moyen de lettres par mot ;

X_2 **:** le nombre moyen de mots par phrase.

Cette formule, tout ce qu'il y a de plus classique, atteint un R de $0,862$, un taux de corrélation multiple égal ou supérieur aux cinq formules de L1 étudiées par cet auteur. Cependant, à la différence de Brown, l'écart paraît ici trop faible pour conclure à la supériorité des formules spécifiques à la L2.

Cherchant à réconcilier ces deux résultats, Greenfield (2004) remarque que la formule de Brown est développée sur la base d'un corpus bien plus hétérogène que ne le sont les 30 passages de Bormuth qu'il a lui-même employés. De plus, en appliquant la formule de Brown à ces 30 extraits, il observe une corrélation de $-0,82$ avec le critère, ce qui est nettement supérieur à ce qu'avait obtenu Brown. Il en déduit que (1) ce sont les particularités du corpus de Brown qui expliquent les corrélations peu élevées obtenues (il aurait été trop difficile pour les sujets impliqués) et (2) que la formule de Brown et la sienne, de même que les formules classiques sont toutes adaptées à l'évaluation de textes académiques, mais fonctionnent nettement moins bien sur un corpus de textes quelconques tel celui utilisé par Brown. Autrement dit, Brown compare les performances des formules de L1 obtenus sur son corpus – ce qui revient à effectuer une validation croisée – avec le score de sa propre formule qui est directement calculé à partir du corpus d'entraînement. Il est donc normal que le second résultat soit supérieur.

Pour autant, une critique peut être faite à Greenfield (2004). Celui-ci compare des formules dont les variables ne sont pas spécifiques à la lecture en L2. Dès lors, il est plus probable qu'il n'observe que de faibles différences entre des formules pour la L1 et la L2. Comme y ont déjà insisté précédemment certains chercheurs (Carrell, 1987; Laroche, 1979), il convient de s'intéresser à des variables tenant compte des spécificités du processus de lecture en L2.

Heilman *et al.* (2007), dont nous avons déjà présenté les travaux pour la L1, vont explorer cette piste. Partant du constat que la plupart des apprenants intègrent les structures syntaxiques de la L2 via un enseignement explicite et progressif, ils suggèrent de prendre en compte 22 de ces motifs comme prédicteurs de la difficulté. Ils considèrent également d'autres facteurs plus aisés à détecter automatiquement, tels que différents temps et modes, ainsi que certaines catégories de discours. Ces variables sont combinées au sein du kNN décrit précédement. L'hypothèse des auteurs semble se vérifier puisqu'ils observent que l'apport des variables syntaxiques au sein du kNN est supérieur pour le modèle en L2 qu'en L1. Cela confirme la pertinence de formules comportant des facteurs propres à la lecture en L2.

Weir et Anagnostou (2008) mettent en valeur un autre fait linguistique susceptible de poser des difficultés particulières à des apprenants lors d'une lecture en L2 : les collocations. Ils proposent une variable qui n'est autre que la fréquence moyenne des collocations du texte à analyser, ladite fréquence ayant été préalablement estimée sur un corpus externe. Ils suggèrent cette fréquence moyenne des collocations comme moyen de mesurer la lisibilité de textes pour l'ALS. Toutefois, leur étude reste incomplète, car elle ne présente ni validation de cette hypothèse sur des données, ni une formule de lisibilité qui en découlerait.

La dernière formule en date pour l'ALS est due à Crossley *et al.* (2008). Les auteurs dénoncent la tendance des travaux en lisibilité des L2 qui se limitent à une approche classique, sans prendre en compte ni les modèles psycholingusitiques de la lecture actuels en L1 et L2, ni les facteurs mis à jour par la nouvelle lisibilité. Grâce à leur outil Coh-Metrix présenté précédemment, ils ouvrent la lisibilité de l'ALS aux techniques de la linguistique computationnelle.

Dans leur étude, ils utilisent, comme critère pour leur formule de lisibilité en L2, les résultats des 200 participants japonais de Greenfield à des tests de closure sur les 32 passages de Bormuth – qui, soulignons-le, commencent à dater (ils remontent à 1971). Ne disposant que de très peu de données, ils se voient obligés de limiter drastiquement le nombre de prédicteurs et en sélectionnent trois au départ de postulats théoriques : un indice de fréquence repris à la base CELEX (Baayen *et al.*, 1993) (X_1) ; la similarité syntaxique entre phrases, à savoir la moyenne, pour chaque phrase adjacente du texte, de la proportion de noeuds en commun entre leurs arbres de dérivation (ce facteur permet d'évaluer la redondance au niveau des structures sémantiques) (X_2) ; un indice de la cohésion textuelle : la proportion de mots de contenu partagés par deux phrases adjacentes, normalisée en fonction du nombre total de mots de contenu (X_3).

À l'aide de cet outillage méthodologique et d'une régression linéaire multiple, les auteurs

conçoivent le *Coh-Metrix L2 reading index* (Crossley *et al.*, 2008, 487) :

$$\text{Predicted cloze} = -45,032 + 22,205 * X_1 + 61,306 * X_2 + 52,23 * X_3 \quad (2.56)$$

Ils rapportent d'excellents résultats : un R de $0,84$ [64] en validation croisée de type *leave-one out* et en concluent que l'addition de variables propres aux aspects cognitifs de la lecture en L2 est tout de même désirable.

En conclusion, il faut noter que les travaux actuels en lisibilité des L2 restent en retard par rapport aux études sur les L1. En effet, seules deux études (Crossley *et al.*, 2008; Heilman *et al.*, 2007) peuvent être rattachées au paradigme de la lisibilité computationelle qui domine aujourd'hui largement la discipline. Cet état de fait s'explique peut-être par le fait que les recherches en lisibilité des L2 requièrent davantage de connaissances en didactique et que les experts en ce domaine sont moins à l'aise avec les techniques du TAL. Cela montre en tout cas que des études telles que la nôtre sont nécessaires et, comme nous allons le découvrir dans la section suivante, c'est encore plus vrai en ce qui concerne le FLE.

2.3.5 Travaux récents en FLE : pas grand chose à se mettre sous la dent !

On ne compte guère plus de trois études récentes sur le FLE, ce qui confirme la nécessité d'effectuer de nouvelles recherches dans ce domaine. La première d'entre-elles est due à Cornaire (1985, 1988), qui se pose une question similaire à celle de Hamsik : sachant que les formules d'Henry (1975) sont les seules disponibles pour le français, est-il possible de les adapter au contexte du FLE ? Plus précisément, ayant décidé – pour des raisons pratiques – de se concentrer sur la formule courte d'Henry (qui comporte les trois variables suivantes : MP, AG et DEXGU (*cf.* Section 2.2.5), Cornaire rencontre le problème suivant. Henry a développé trois formules courtes, adaptées respectivement à des élèves de la 5e et 6e primaire, de la 2e et 3e secondaire et de la 5e et 6e secondaire. Comment faire correspondre cette échelle scolaire aux niveaux d'apprenants en FLE ?

Pour répondre à cette question, Cornaire va dans un premier temps vérifier la validité prédictive de la formule d'Henry pour le contexte québécois en L1. Elle compare, sur trois textes de presse, les prédictions de la formule avec les résultats d'étudiants à des tests de closure. Ces étudiants sont issus des niveaux secondaires I, III et V (ssytème québecois), qui d'après Cornaire correspondent aux niveaux envisagés par Henry (1975). Elle note que si pour les années III et V, les résultats du test de closure et de la formule sont relativement cohérents, la formule semble sous-estimer les performances des élèves de première secondaire. Ceci

64. Il faut cependant rappeler que les passages de Bormuth sont connus pour générer des corrélations élevées, comme nous l'avons déjà mentionné auparavant.

pourrait toutefois s'expliquer par un effet de motivation des sujets envers l'étude, estime Cornaire (1988).

Au terme de cette première expérience, elle dispose d'un score de closure moyen pour chaque texte et chaque niveau scolaire. Dès lors, soumettant un échantillon d'apprenants de FLE au même test, il lui est possible d'associer chacune des trois formules à un public cible pour le FLE. Cette seconde expérimentation révèle une correspondance entre, d'une part, les apprenants intermédiaires avancés et les élèves du secondaire I et, d'autre part, les apprenants avancés et les étudiants du secondaire III. Il serait donc justifié d'appliquer une formule de L1 à un contexte de L2. Cornaire reste toutefois réservée sur la validité de ses résultats et appelle à répliquer ce type d'expériences.

Ce vœu ne sera malheureusement pas réalisé, puisque les recherches dans le domaine restent en suspens jusqu'aux travaux de Uitdenbogerd, qui endosse une optique différente. Dans un premier article, Uitdenbogerd (2003) doute que les formules développées pour des enfants lisant dans leur langue maternelle soient adaptées à des apprenants adultes d'une L2 [65]. Elle se propose d'étudier la relation particulière qu'entretiennent une série de facteurs avec la difficulté de textes évaluée par des juges humains. Toutefois, les quatre facteurs qu'elle expérimente en 2003 restent convenus. Il s'agit du nombre de mots par phrase, du nombre de mots de plus de deux syllabes, de facteurs visuels et du nombre de mots que la chercheuse ne peut définir elle-même dans un texte. Par conséquent, l'étude n'est pas conclusive sur la nécessité d'une formule particulière au contexte L2, pas plus qu'elle ne propose de formule concrète. Tout au plus défend-elle l'importance de la syntaxe comme facteur en lisibilité du FLE.

Dans un article postérieur, Uitdenbogerd (2005) va plus loin dans ses expérimentations et développe une formule spécifique au FLE qui prend en compte les congénères français pour un apprenant anglophone. Pour son corpus, elle recourt à l'aide de quinze apprenants de FLE qui ordonnent dix livres en fonction de leur complexité. Elle en dérive l'équation suivante dont les coefficients ont été définis manuellement :

$$\text{FR} = 10 * WpS - Cog \tag{2.57}$$

dans laquelle

WpS **:** le nombre moyen de mots par phrase ;

Cog **:** le nombre de congénères pour 100 mots.

On notera la présence de certains problèmes méthodologiques. Sans compter la taille réduite

65. Notons que si cette auteure australienne est l'une des rares à s'attacher au FLE, elle se situe résolument dans une perspective anglo-saxonne. En effet, les formules en L1 qu'elle envisage sont celles de Flesch (1948) ou Dale et Chall (1948a). Elle ne fait aucune référence à des travaux sur le français dans cet article.

du corpus, le critère utilisé nous semble quelque peu manquer de rigueur : la difficulté des livres est évaluée partir des 100 premiers mots, alors que les incipits sont connus pour être moins représentatifs de la difficulté d'un ouvrage ; les méthodes employées par les apprenants ne sont pas nécessairement similaires. Par ailleurs, on regrette aussi l'absence d'une mesure lexicale classique, qui, au-delà des congénères, prenne en considération les variations au niveau du vocabulaire. Malgré tout, l'étude rapporte un résultat particulièrement intéressant. Cette nouvelle mesure obtient de meilleures performances pour les apprenants du FLE que des formules traditionnelles pour l'anglais. Par contre, elle se comporte moins bien pour les deux natifs, ce qui laisse penser que la conception de formules spécifiques au contexte L2 et la prise en compte des congénères sont justifiées.

Les derniers travaux publiés sur le FLE, à notre connaissance, sont notre contribution (François, 2009a,b). Ils se situent dans le prolongement de la lisibilité computationnelle et se focalisent sur des variables de surface. En nous basant sur un corpus de manuels de FLE, nous avons pu proposer plusieurs formules dont la difficulté est exprimée selon une échelle particulièrement pratique, celle définie dans le CECR. Nous ne détaillerons pas ici les détails de la conception de ce nouvel outil, puisque cela fera l'objet des prochains chapitres de cette thèse.

2.3.6 Conclusion

De ce parcours sur les études de lisibilité en L2, il nous semble que se dégagent trois problématiques majeures : le manque d'études dans le domaine, la question de la nécessité d'une formule spécifique aux L2 et la question de l'emploi des variables structuro-cognitives. Si nous avons déjà insisté sur la première d'entre elles, qui nous chagrine d'autant plus que le domaine offre pourtant de nombreuses opportunités pédagogiques (*cf.* Chapitre 1), il est intéressant de revenir sur les deux autres.

En ce qui concerne la nécessité de formules particulières, nous avons montré qu'il existe deux opinions opposées. Si Hamsik (1984, cité par Greenfield 2004), Greenfield (1999, cité par Greenfield 2004) et Cornaire (1985) ont confirmé la validité de formules pour natifs lorsqu'on les applique à un contexte de L2, Brown (1998, cité par Crossley *et al.* 2008), Uitdenbogerd (2005) et François (2009a) mettent en question la possibilité d'un tel transfert.

À la lecture de ces différents travaux, il nous semble qu'on peut en dégager les trois constats suivants : (1) les formules de lisibilité développées pour la L1 restent valides dans le contexte des langues étrangères, mais (2) il arrive que les formules calculées sur des données spécifiques à cette seconde population fournissent de meilleures performances. Il serait donc justifié de concevoir des outils adaptés à ce contexte particulier, d'autant que (3) divers

chercheurs (François, 2009a; Laroche, 1979; Uitdenbogerd, 2005) ont mis en évidence l'intérêt de prendre en compte certaines variables, telles que la proportion de cognates, n'ayant de sens que dans un contexte L2.

En ce qui concerne l'emploi des variables structuro-cognitives, le domaine apparaît largement en retard sur les études en L1. Seuls Crossley *et al.* (2008) ont testé une formule intégrant un facteur de ce type et rapportent un effet significatif pour celui-ci. Ce moindre attrait pour les indices de niveaux supérieurs s'explique probablement par une bonne raison théorique. Nous avons vu précédemment que les mesures de lisibilité de l'approche cognitivo-structurale sont plus discriminantes pour du matériel de haut niveau, dans lequel c'est le contenu qui pose problème. Or, en L2, le décodage constitue plus longtemps une source majeure de difficulté, puisque les structures de la langue ne préexistent généralement pas à l'apprentissage de la lecture. Cela pourrait expliquer pourquoi la syntaxe et le lexique restent les facteurs dominants dans le domaine.

Pour autant, il nous semble que ces deux questions restent pertinentes et n'ont pas fait l'objet de suffisamment d'études. Nos propres travaux ont pour but d'apporter quelques pierres supplémentaires à l'édifice de la lisibilité dans le contexte des L2.

2.4 Conclusion générale

Dans ce second chapitre, nous avons parcouru 90 ans de recherches en lisibilité, depuis les débuts de la discipline avec Lively et Pressey (1923), jusqu'aux travaux les plus récents (Kate *et al.*, 2010; Tanaka-Ishii *et al.*, 2010). Nous avons montré que si les méthodologies appliquées ont fortement évolué depuis les origines, certaines questions restent ouvertes : quelles sont les variables les plus efficaces pour évaluer la lisibilité de nouveaux textes ? Quel est le meilleur moyen de mesurer la compréhension des lecteurs ? Pour des raisons d'organisation du contenu, nous reviendrons sur ces deux aspects dans les chapitres 3 et 4, qui traitent plus particulièrement des aspects méthodologiques de la conception d'une formule de lisibilité en FLE.

Nous avons également insisté sur le fait que la lisibilité doit aujourd'hui être envisagée comme un modèle prédictif qui, pour une population aussi homogène que possible, peut poser un diagnostic sur un ensemble de textes dont la difficulté varie. À l'aide de cette définition, il nous semble possible d'éclairer autrement le grand débat des années 70-80. En effet, si la lisibilité classique a produit des formules pratiques, elle a aussi été envisagée par certains comme une première tentative d'explication de la compréhension des textes, au même titre que les modèles de lecture qui naissent ensuite avec la psychologie cognitive. À l'instar de ceux-ci, il s'agissait de mesurer la compréhension qu'une population déterminée

de sujets avait d'un ensemble de textes pour ensuite chercher à expliquer la variation de cette variable dépendante.

Le problème est que ces techniques s'appuyaient exclusivement sur des caractéristiques textuelles (toutefois passées au travers du prisme du lecteur grâce au test de compréhension). Avec raison, elles se sont vues critiquées en tant que modèle de la lecture, en ce qu'elles ne tenaient compte ni des variables contextuelles, ni surtout de dimensions textuelles précédemment inconnues. Dès lors, il nous apparaît que les critiques structuro-cognitivistes, plutôt que de détruire l'apport classique, ont recadré les ambitions de la lisibilité, tout en ouvrant de nouvelles perspectives quant aux variables prises en compte.

Sur la base de cette hypothèse, il est possible de délimiter trois grandes mouvances en lisibilité : la lisibilité classique, la lisibilité structuro-cognitive et ce nouvel élan que nous avons dénommé la lisibilité computationnelle. Nous y retrouvons un schéma classique de thèse, antithèse et synthèse. La lisibilité classique a posé une série de bases méthodologiques qui ont donné naissance à des modèles employés en éducation. Elle a ensuite subi une critique radicale, dont les partisans n'ont cependant pas abouti à des meilleures formules. Enfin, les avancées en TAL ont permis d'envisager la conception de formules qui combinent l'ensemble des facteurs découverts par les deux tendances précédentes dans un cadre automatisé.

C'est cette troisième voie qui nous semble aujourd'hui la plus prometteuse dans le domaine. En effet, elle laisse présager l'intégration des apports classiques et structuro-cognitives au sein des modèles prédictifs qui prendraient en compte l'ensemble des dimensions textuelles actuellement reconnues. Bien sûr, pour prendre correctement en compte ces aspects sémantiques et cognitifs, il est encore nécessaire d'effectuer certaines avancées dans les techniques de TAL, en particulier en ce qui concerne le français. Mais les outils actuels ouvrent déjà de nombreuses pistes qui n'ont guère encore été explorées, en particulier en FLE.

Enfin, il faut souligner que puisque les études sur les causes réelles des difficultés de lecture ont également bien progressé en psychologie cognitive, il n'est désormais plus exclu d'envisager des formules de lisibilité où chacun des facteurs ne serait pas seulement un indice, c'est-à-dire une dimension corrélée avec la difficulté, mais bien une cause réelle de celle-ci. Malgré cela, les concepteurs de formules de lisibilité ne doivent pas oublier que leur objectif principal reste la prédiction et qu'il convient, par conséquent, de préférer les facteurs les plus efficaces, quand bien même ceux-ci ne constitueraient pas des causes reconnues de la difficulté. C'est pourquoi nous avons mêlé, dans notre étude, ces deux types de facteurs, qui sont décrits dans le chapitre suivant.

Deuxième partie

Méthodologie : un modèle pour la mesure de la difficulté

Introduction

Les étapes méthodologiques pour la conception d'une formule de lisibilité

La première partie de cette étude a passé en revue les travaux traitant des processus de la lecture et des formules de lisibilité. Dans la seconde partie, nous décrivons plus précisément comment ce dépouillement a orienté nos choix méthodologiques dans nos expérimentations sur les apports du TAL à la lisibilité et dans la conception d'une formule pour le FLE. Nous allons détailler chacune des grandes étapes de la conception d'une formule et critiquer les différentes positions méthodologiques communes dans la littérature, avant de justifier le ou les choix que nous avons opérés parmi celles-ci.

Ces grandes étapes méthodologiques pour la conception d'une formule de lisibilité sont les suivantes :

1. Constituer un corpus d'entraînement annoté. Par le terme « annoté », nous indiquons que la difficulté de chaque texte à la lecture a été mesurée à l'aide d'un critère. Cette première étape soulève trois questions principales : quels textes faut-il retenir au sein du corpus d'entraînement ; quel critère est le plus adéquat pour mesurer la compréhension de lecteur en L2 ; quelle échelle de difficulté utiliser ? Nous avons déjà montré l'importance de ces questions en lisibilité et introduit les débats qui les entourent dans le Chapitre 2. Cependant, le Chapitre 4 détaillera ces trois problématiques en se concentrant sur les aspects qui sont pertinents pour notre problématique de recherche.
2. Parallèlement à cette étape de collecte des données, il faut aussi identifier divers facteurs textuels susceptibles de rendre compte des différences de difficulté entre les différents textes du corpus. Certes, nous avons déjà abordé de nombreux aspects de cette problématique au cours de notre parcours chronologique (*cf.* Chapitre 2). Toutefois, le Chapitre 3 passera en revue les variables qui apparaissent utiles pour la lisibilité du FLE, en insistant sur les fondements théoriques ou empiriques de chacune d'elles. Nous expliciterons également les choix techniques opérés pour paramétriser les variables que nous avons retenues pour la partie expérimentale de notre étude.
3. Enfin, il convient de définir un modèle qui sélectionne les caractéristiques textuelles les plus informatives et apprend automatiquement les poids à leur attribuer sur la base

du corpus de textes. Divers algorithmes statistiques ont été envisagés dans ce but. Leur principe est résumé rapidement dans le Chapitre 5, mais nous n'avons pas dédié un chapitre spécifique à leur description étant donné qu'il s'agit de techniques standards, bien documentées par ailleurs.

En résumé, cette seconde partie de notre thèse se situe donc entre la théorie et la pratique. Les notions qui y sont discutées sont essentiellement théoriques, mais leur présentation et la critique qui en est faite sont orientées par nos objectifs de recherche : concevoir une formule de lisibilité spécifique à la lecture de textes en FLE, tout en posant un regard critique sur la lisibilité computationnelle.

Quelles hypothèses pour une modélisation de la difficulté des textes ?

Avant d'entrer dans le vif du sujet, il nous paraît instructif de s'arrêter un instant pour réfléchir aux hypothèses, parfois assez fortes, qu'il est nécessaire d'émettre lorsqu'on modélise la difficulté des textes au moyen d'une formule de lisibilité. En effet, on ne peut appliquer n'importe quelle formule à n'importe quel contexte, comme l'avaient rappelé certains des structuro-cogitivistes (*cf.* Chapitre 2.2.6). Bruce *et al.* (1981, 9-10), en particulier, avaient d'ailleurs raillé l'usage de formules en listant une série de conditions théoriques qui ne sont jamais respectées.

Afin d'éviter de tomber dans cet écueil, mais surtout afin d'avertir le lecteur sur les limites de notre propre modèle, nous avons rassemblé ici les principaux postulats sur lesquels repose la conception et l'utilisation de notre formule de lisibilité pour le FLE.

Hypothèse 2.1 *Il est possible de mesurer quantitativement la compréhension d'un texte particulier chez un lecteur donné. Autrement dit, il existe au moins un dispositif qui produit, pour un texte et un lecteur donnés, une valeur unique qui représente bien une mesure de la compréhension du texte par cet individu.*

Hypothèse 2.2 *En conséquence de l'hypothèse précédente, on postule qu'il est possible de mesurer la compréhension moyenne qu'une population spécifique de lecteurs se fait d'un texte (ou, par extension, d'un ensemble de textes), en se basant sur des mesures prises sur un échantillon représentatif.*

Hypothèse 2.3 *La lecture pratiquée par les lecteurs de cet échantillon correspond à celle qui est enseignée dans une classe de FLE utilisant une approche communicationnelle. Elle peut dès lors cibler différents objectifs : extraire une information (balayage), uniquement identifier les idées principales (écrémage) ou comprendre de manière plus détaillée le contenu véhiculé dans les textes du corpus.*

Hypothèse 2.4 *Pour prédire la difficulté de compréhension des textes du corpus telle qu'elle a été mesurée par le dispositif décrit ci-dessus, des* ***indices*** *qui reflètent les difficultés de lecture suffisent. Il n'est pas nécessaire que les variables explicatives considérées constituent des causes attestées des difficultés de lecture.*

Nous avons vu que cette hypothèse a été remise en question par les chercheurs de la nouvelle lisibilité (*cf.* Chapitre 2.2.6), qui ont pourtant finalement dû accepter que les facteurs causaux ne représentent pas de meilleurs prédicteurs que les simples indices. C'est pourquoi nous considérons que cette hypothèse reste valide. D'autant que, derrière cette opposition, il nous semble qu'on retrouve la distinction bien connue en statistique entre un modèle prédictif et un modèle explicatif. Les formules de lisibilité appartenant plutôt aux modèles de type prédictif, elles n'ont pas besoin de déterminer précisément les causes du phénomène à prédire, pour autant qu'elles aient identifié des prédicteurs efficaces.

Hypothèse 2.5 *Une approche de la lisibilité des textes qui adopte une position explicative et, par conséquent expérimentale, est la « writability ». Dans ce cas, les facteurs manipulés doivent être des causes effectives de la difficulté des textes, puisqu'il s'agit de réécrire des textes afin de modifier leur niveau de difficulté. C'est pourquoi, si cette condition n'est pas respectée, il est dangereux d'appliquer telle quelle une formule de lisibilité, c'est-à-dire un modèle prédictif, en « writability ».*

Hypothèse 2.6 *Si l'on considère qu'une même formule suffit pour une population spécifique des lecteurs, quels que soient leur niveau, la situation de lecture, etc., on doit postuler que l'importance des variables est constante quel que soit le niveau des lecteurs.*

Cette dernière hypothèse n'est pas acceptée par l'ensemble des auteurs. Si Bormuth (1966) était arrivé à une telle conclusion, d'autres auteurs tels qu'Henry (1975) n'ont pas été convaincus et ont proposé des formules différentes en fonction du niveau des apprenants.

Hypothèse 2.7 *La relation des indices de difficulté à la variable dépendante mesurée par le dispositif n'est pas nécessairement linéaire.*

Cette constatation, faite en premier par Bormuth (1966), nous a conduit à recourir à d'autres modèles statistiques que la régression linéaire classiquement utilisée en lisibilité.

C'est avec cette dernière hypothèse que nous clôturons cette introduction à la partie méthodologique. Comme nous l'avons annoncé précédemment, les deux prochains chapitres vont couvrir deux des trois étapes principales dans la constitution d'une formule de lisibilité : la conception du corpus et la définition de prédicteurs.

CHAPITRE 3

QUELLES VARIABLES POUR UN MODÈLE EN FLE ?

3.1 Introduction

> *« Without question the most important advances in readability research should result from the development of new linguistic variables. » Bormuth (1966, 86)*

Cette citation de Bormuth (1966) nous rappelle une évidence : en lisibilité, l'étude des aspects textuels susceptibles d'entretenir au moins un lien de corrélation avec la difficulté des textes est capitale. C'est pourquoi nous consacrons à cette problématique le premier chapitre de la partie méthodologique de notre thèse. Notre objectif est double : il s'agit d'abord de décrire l'ensemble des dimensions textuelles qui ont été prises en compte dans la littérature ou qui gagneraient à l'être, avant de détailler la sélection que nous avons opérée parmi celles-ci.

Afin de réaliser le premier de ces deux objectifs, nous avons recouru à deux stratégies différentes. D'une part, nous avons compulsé l'importante littérature en lisibilité de l'anglais et du français L1, ainsi que les quelques travaux portant sur la L2, ce qui nous a permis d'identifier un grand nombre de facteurs de difficulté plus ou moins connus. D'autre part, nous avons exploré diverses études en psychologie cognitive de la lecture et en neurosciences afin d'y découvrir des caractéristiques textuelles qui ralentissent ou facilitent le processus de lecture et qui ont peu ou prou fait l'objet d'études dans le cadre de la lisibilité.

Ainsi, la première partie de ce chapitre décrit les différents prédicteurs à notre disposition au terme de cette double exploration théorique. Dans la deuxième partie, nous adoptons un point de vue pratique où seules les variables retenues dans notre modèle se retrouvent. Nous explicitons de manière plus formelle la manière dont nous avons paramétrisé chacune d'elles et justifions les choix opérés lorsque plusieurs alternatives existent.

3.2 Quelles variables pour un modèle de lisibilité en FLE ?

Comme nous l'avons rappelé dans le Chapitre 2, de nombreuses dimensions textuelles ont été étudiées dans le cadre de la lisibilité en L1. Il suffit de se remémorer l'étude exhaustive de Gray et Leary (1935), qui classent pas moins de 289 variables en 4 catégories : le contenu, le style, le format et l'organisation. En ce qui concerne le FLE, nous n'avons pas connaissance d'une approche descriptive aussi exhaustive. Certes, la majorité des prédicteurs considérés pour la L1 restent pertinents en L2. Néanmoins, il en est certains qui nécessitent d'être adaptés ou repensés en fonction de ce contexte particulier. De plus, les dimensions spécifiques au processus de lecture en L2, mises en évidence par divers travaux de psycholinguistique, de psychologie cognitive et par les neurosciences, méritent aussi d'être considérées en lisibilité des L2.

Dans cette première partie de notre réflexion sur les variables adaptées au contexte L2, nous synthétisons le résultat de nos recherches au sein de la littérature en lisibilité et des études sur les processus de lecture en L1 et L2. Nous décrivons un vaste ensemble de prédicteurs regroupés au sein de cinq catégories : les indices lexicaux, les indices syntaxiques, les variables de la difficulté conceptuelle, celles qui mesurent la cohérence textuelle et, enfin, celles qui sont spécifiques au FLE. Pour chacun des prédicteurs considérés, nous rapportons des observations théoriques ainsi que des résultats empiriques qui visent à évaluer leur pertinence. De plus, lorsque c'est justifié, ces remarques sont mises en relation avec notre contexte d'intérêt : le FLE.

Il faut noter que si nous n'hésitons pas à replacer certaines variables dans leur contexte historique, notre objectif n'est pas tant de proposer une liste exhaustive des prédicteurs utilisés dans la littérature que d'établir des fondements théoriques pour notre propre étude, mais aussi pour les études à venir. En effet, il n'a pas été possible d'expérimenter l'ensemble des variables répertoriées ici et ceci pour deux raisons principales. D'une part, certaines d'entre elles requièrent l'emploi de corpus particuliers – par exemple, des textes dont la difficulté a été évaluée pour des apprenants d'une langue maternelle donnée – dont nous ne disposions pas. D'autre part, d'autres variables nécessitent le recours à des traitements de TAL avancés pour lesquels il n'existe pas nécessairement un outil efficace et disponible pour le français, à la différence de l'anglais.

3.2.1 Caractéristiques d'un bon prédicteur en lisibilité

Avant de nous lancer dans la description des prédicteurs, il est utile de s'arrêter un instant sur la problématique suivante. Qu'est-ce qui fait un bon indice de la difficulté des textes ? En parcourant la littérature, nous avons identifié les quatre propriétés suivantes :

1. Un prédicteur efficace doit entretenir la corrélation la plus élevée possible avec le critère choisi pour mesurer la difficulté des textes du corpus d'entraînement. Un taux de corrélation élevé constitue certainement la qualité la plus évidente pour une variable donnée. C'est pourquoi cet outil a été utilisé par nombre de chercheurs comme un moyen simple d'établir l'importance des facteurs. Il faut toutefois noter que plusieurs auteurs (Carrell, 1987; Henry, 1980; Klare, 1984) ont montré qu'un corpus plus diversifié conduisait à un meilleur taux de corrélation. Carrell (1987, 25) explique bien ce phénomène :

   > *If the criterion consists of a large number of tests passages of varying content and with a wide range of difficulty and if a wide range of readers is added, then correlations are increased. However, the predictability of the formulas – that is, the high correlations – drops dramatically when more limited ranges of reader abilities, text subjects and number of texts passages are considered.*

   Il faut donc relativiser les résultats impressionnants de certaines formules, comme celle de Bormuth (1969a) qui a typiquement été entraînée sur un corpus favorable.

2. Dans le même ordre d'idée, un prédicteur efficace doit aussi être le moins corrélé possible avec les autres variables indépendantes, sans quoi il n'apporte que peu ou prou d'informations nouvelles à la formule de lisibilité.

3. Il faut aussi que la variable considérée soit mesurable d'une manière objective et reproductible. C'est ce que Harrisson (1980, 55-56) appelle la fiabilité de l'analyste. C'est-à-dire qu'il faut non seulement être capable de lister l'ensemble des valeurs possibles de cette variable, mais aussi pouvoir définir des règles de classification non ambiguës afin que toute observation soit associable à une valeur unique. Or, comme l'a montré le chapitre 1, certaines dimensions du processus de lecture restent difficilement explicitables et, par voie de fait, difficilement formalisables. On notera que les difficultés de paramétrisation d'une variable varient selon que le comptage se déroule manuellement ou automatiquement.

4. Enfin, parmi ces propriétés objectivement mesurables, il faut encore se limiter à celles qui se répètent un nombre de fois suffisant pour que l'information qu'ils apportent présente un intérêt pour un modèle. Ainsi, on s'accordera probablement à considérer qu'une variable « le texte est écrit par Proust » constitue un sérieux indice de la difficulté d'un texte lorsqu'elle prend la valeur « oui ». Toutefois, ce cas étant très rare dans le contexte d'une classe de FLE, l'information apportée par cette variable sera trop rarement utile pour que ce prédicteur présente un intérêt quelconque au sein d'une formule de lisibilité.

En sus de ces quatre qualités largement acceptées dans la littérature, Bormuth (1969a) identifie deux caractéristiques supplémentaires qu'il estime souhaitable pour les indices de lisibilité. D'une part, il distingue les facteurs qui ont une validité théorique de ceux qui ont uniquement fait leurs preuves empiriquement. Si les premiers chercheurs en lisibilité ne s'inquiétaient pas trop de cette question, Bormuth (1969a, 10) recommande de n'utiliser que des variables qui sont supportées par des arguments théoriques raisonnables, autrement dit qui entretiennent probablement une relation causale avec la difficulté. D'autre part, il fait aussi la différence entre variable manipulable et non manipulable, la première étant évidemment préférable. Dans le cadre de cette étude, nous avons, autant que faire se peut, toujours cherché à rassembler les arguments théoriques qui soutiennent, ou infirment, les prédicteurs que nous envisageons. Toutefois, comme il nous semble que ces deux qualités trouvent davantage de pertinence dans le cadre d'un modèle explicatif que celui d'un modèle prédictif, nous nous sommes refusés à exclure des variables, non fondées théoriquement, mais que soutiennent les expériences de nombreux chercheurs au fil des années.

3.2.2 Les variables au niveau lexical

La première catégorie de facteurs que nous envisageons concerne les aspects lexicaux des textes et cherche à modéliser les difficultés qui naissent du lexique. L'importance de ce type d'indices de lisibilité ne fait aucun doute : ils ont souvent été considérés comme primordiaux pour la compréhension à la lecture (Chall et Dale, 1995, 82). Il s'agit aussi des variables ayant fait l'objet des premières études, comme nous l'avons rappelé dans le Chapitre 2. Aujourd'hui, on sait que ces facteurs agissent essentiellement sur les processus de reconnaissance des mots, qui participent au décodage du texte. Nous avons montré, dans le Chapitre 1, que cette étape est essentielle pour la lecture, et plus particulièrement la lecture en L2, puisque des problèmes de décodage peuvent engendrer des difficultés de compréhension.

Dans cette section, nous allons passer en revue la majorité des variables lexicales qui ont été utilisées en lisibilité, tout en proposant quelques perspectives neuves qui s'inspirent des connaissances en psychologie cognitive de la lecture. Comme Chall (1958, 45) l'avait déjà signalé, « toutes les mesures de la difficulté du vocabulaire sont corrélées ». Dès lors, notre objectif est ici de procéder à un double démêlage théorique : non seulement, en distinguant clairement chaque famille de variables des autres, mais également en cherchant à déterminer l'existence d'un effet spécifique à chacune d'elle, c'est-à-dire un effet sur la compréhension à la lecture qui ne soit pas dû à une autre variable corrélée avec notre facteur d'intérêt. Ainsi, nous détaillerons ainsi successivement des mesures basées sur la fréquence lexicale, la vraisemblance textuelle, la familiarité, la diversité lexicale, l'âge d'acquisition, le voisinage

orthographique et la longueur des mots.

Mesures de la fréquence lexicale

Cette première famille de mesures de la complexité lexicale repose sur une même hypothèse :

Hypothèse 3.1 *Il existe une association forte entre la fréquence d'emploi des mots dans la langue (ou fréquence objective) et leur facilité de leur lecture, de sorte qu'il est possible de se servir de la première pour prédire la seconde. Autrement dit, plus un texte comporte des mots rares, plus il peut être considéré comme lexicalement ardu.*

Justifications empiriques et théoriques : Depuis le début des recherches en lisibilité, cette association entre fréquence objective et facilité lexicale s'est toujours révélée fructueuse, comme le rappelle Henry (1975, 74) : « quelle que soit la liste de fréquence utilisée, on a toujours trouvé une relation importante entre le poids du vocabulaire mesuré en termes de fréquence et la difficulté du texte évalué selon un critère externe ».

Pourtant, cette hypothèse est longtemps restée peu fondée théoriquement, s'appuyant essentiellement sur des justifications empiriques ou des réflexions basées sur le sens commun. On ne peut résister à l'envie de donner en exemple les propos de Tharp (1939, 2) qui défend cette association par l'analogie :

> *The resistance, load, charge, "burden" of a reading unit, met many times in a given spread of running words, is less than one met rarely. You encounter Fred Smith five times every day, Peter Jones once every two weeks : which do you recognize most readily in a crowd of people ?*

Plusieurs décennies après, (Klare, 1968, 15) étudie cette question dans un article où il dit :

> *Frequency of occurrence of words, as this paper indicates clearly, plays an all-pervasive role in language usage. Not only do humans tend to use some words more often than others, they recognize more frequent words more rapidly than less frequent, prefer them and understand and learn them more readily. It is not surprising, therefore, that this variable has such a central role in the measurement of readability.*

Cependant, c'est d'études psychologiques sur les facteurs influençant la reconnaissance des mots que proviennent des preuves solides en faveur de cette association. L'article séminal est dû à Howes et Solomon (1951) qui étudient la relation entre la probabilité de mots et la

vitesse de reconnaissance. Ils mettent à jour une corrélation linéaire négative (de $-0,68$ à $-0,75$)[1] entre le logarithme de la fréquence des mots (calculées selon trois sources différentes) et le temps d'exposition minimum nécessaire à leur reconnaissance.

Par la suite, de nombreux chercheurs s'accordent à dire que la fréquence des mots influe sur les tâches de décision lexicale, d'identification perceptive, de prononciation, de catégorisation sémantique, mais également sur les mouvements des yeux[2]. Cet effet, connu sous le terme d'effet de fréquence (Lupker, 2005, 42), a été étudié, entre autres, par Becker (1976), Monsell (1991), Inhoff et Rayner (1986) ou encore, en ce qui concerne le français, par O'Regan et Jacobs (1992). Brysbaert *et al.* (2000, 66) expliquent que cet effet se justifie parce que « les représentations des mots communs dans le lexique mental sont plus facilement accessibles que celles de mots moins communs ».

Au-delà de la simple reconnaissance de cet effet de fréquence, divers travaux poussent plus loin la compréhension de ce phénomène. Les expériences de Solomon et Postman (1952) montrent que le gain de vitesse dans la reconnaissance est présent, même quand le sens des mots n'est pas compris. Ce serait donc la familiarité avec la forme qui compte. Ce résultat est confirmé par Miller et Kintsch (1980, 384) qui avancent l'hypothèse suivante :

> *The commonly cited variables of word frequency and sentence length appear to affect reading time but no recall. Once a text segment has been decoded, word frequency no longer matters, presumably because further processing occurs on a conceptual level that is unaffected by these factors.*

Autrement dit, ce serait au niveau du décodage des formes lexicales que la fréquence des mots influence la vitesse de lecture. Toutefois, comme le rappellent (Crossley *et al.*, 2008, 488), un décodage mieux automatisé permet de mieux comprendre un texte, puisque la reconnaissance des mots mobilise dès lors moins de ressources mémorielles, lesquelles peuvent être allouées à des processus de niveaux supérieurs relatifs à la compréhension. Cela explique qu'indirectement la fréquence a un effet sur le taux de compréhension d'un texte.

Critiques : Malgré un succès jamais démenti et des assises théoriques solides, cette hypothèse a également fait l'objet de plusieurs critiques et ce, dès les débuts de la lisibilité. Ainsi, Ojemann (1934) et Gray et Leary (1935) mettent en évidence le fait que la difficulté d'une même forme lexicale peut varier considérablement selon le sens activé par le contexte. On peut toutefois répondre à cette objection en calculant les fréquences des sens plutôt que

1. Par la suite, ils corrigent un biais provenant de certaines caractéristiques des lettres composant les mots à reconnaître. Dès lors, les corrélations s'élèvent pour atteindre de $-0,76$ à $-0,83$.

2. Nous renvoyons le lecteur à la synthèse de Ferrand (2007, 91) qui liste plus d'une quarantaine de publications portant sur ces cinq tâches.

celles des formes lexicales. C'est d'ailleurs ce que Lorge et Thorndike ont fait au sein de leur liste *A Semantic Count of English Words*, publiée en 1938.

D'autres auteurs, tels que Morrison et Ellis (1995), remettent en question les effets de la fréquence, estimant que la majorité d'entre eux sont des artefacts de l'âge d'acquisition des mots, une variable hautement corrélée avec la fréquence lexicale. Cette hypothèse a toutefois été infirmée par plusieurs études postérieures. Ainsi, Brysbaert (1996) rapporte, pour le néerlandais, un effet de la fréquence lexicale une fois que l'âge d'acquisition est contrôlé. Peu après, Brysbaert *et al.* (2000) effectuent des expériences supplémentaires et parviennent à la conclusion qu'aussi bien la fréquence d'occurrence que l'âge d'acquisition influencent le temps de latence lors d'une tâche de prononciation ou de décision lexicale en néerlandais. Par contre, il semblerait que ces deux variables n'agissent pas au même niveau. Pour Gerhand et Barry (1998), l'effet de la fréquence jouerait plutôt au niveau de la reconnaissance de mots, tandis que l'âge d'acquisition affecterait davantage les tâches phonologiques.

Une troisième incertitude concernant l'effet de fréquence porte sur son rapport à la longueur des mots. Comme Zipf (1965) l'a démontré, plus un mot est utilisé fréquemment et plus il tend à être raccourci. Dès lors, en ce qui concerne les mouvements oculaires, du moins, il est difficile de déterminer si un temps de fixation plus long provient de l'effet de fréquence ou de la longueur des mots. Toutefois, les études de Inhoff et Rayner (1986) et Rayner et Duffy (1986), qui ont contrôlé la longueur des mots, ont montré que l'effet de fréquence reste bien présent.

En conclusion, on peut affirmer que les fréquences lexicales, bien que corrélées avec d'autres effets lexicaux tels que la familiarité ou la longueur des mots, ont effectivement un effet sur la vitesse de décodage des mots, et partant, sur la compréhension elle-même. Dès lors, il apparaît essentiel d'utiliser une mesure de la fréquence lexicale dans une formule de lisibilité.

La fréquence pour les lecteurs en L2 : L'effet des fréquences lexicales mérite d'être nuancé dans le cas des lecteurs dans une L2. A priori, on s'attendrait à ce qu'il soit encore plus important, puisque c'est au niveau du décodage des formes lexicales que la fréquence des mots influencerait la lecture et que cette étape est particulièrement importante dans le cas de la lecture en L2 tant que les apprenants n'ont pas atteint un certain seuil de compétence. Cependant, Laroche (1979, 132) souligne judicieusement que si, pour un locuteur natif, la difficulté relative d'un mot à la lecture est apparentée à la fréquence de ce mot dans le bain de langue (écrit et oral) dans lequel il est immergé depuis son enfance, ce concept n'a pas de sens pour un apprenant d'une L2.

En effet, ce qui importe n'est pas tant la fréquence des mots dans la langue que la fréquence

à laquelle l'apprenant rencontre ces mêmes mots dans sa pratique de la L2. Autrement dit, il s'agit de sa familiarité avec le vocabulaire de cette langue, une variable dont nous reparlerons dans la suite de ce chapitre. Cependant, si cette objection semble invalider la fréquence objective des mots comme **cause** de la complexité lexicale dans une L2, il n'en reste pas moins que cette information peut constituer un **prédicteur** efficace. En effet, nous avons rapporté plus haut que la fréquence lexicale était largement corrélée à l'âge d'acquisition et à la longueur des mots.

De plus, il faut tenir compte d'un autre phénomène qui milite en faveur de l'insertion de mesures des fréquences lexicales dans la L1 au sein de formules de lisibilité pour la L2. Pour les apprenants qui entrent essentiellement en contact avec des matériaux linguistiques via des manuels, il est clair que leur familiarité avec les mots de la L2 étudiée est régie par la progression établie par les ressources pédagogiques qu'ils utilisent. Or, cette progression au niveau de l'apprentissage du vocabulaire repose fort probablement sur des listes de fréquences, telles que le Français Fondamental de Gougenheim *et al.* (1964), elles-mêmes établies à partir des fréquences en L1. La boucle est bouclée et on en revient à penser qu'une mesure de la fréquence lexicale a sa place dans une formule de lisibilité pour le FLE.

Problématiques particulières : Évaluer la complexité lexicale au moyen de la fréquence des mots soulève un certain nombre de problématiques que nous présentons ici dans leur dimension théorique. Les stratégies que nous avons utilisées pour répondre concrètement à ces problèmes sont envisagées dans la seconde partie de ce chapitre.

Le premier de ces obstacles est aussi le plus évident. Comment estimer, le plus précisément possible, la probabilité de chaque mot qu'il est possible de rencontrer dans un texte ? Il s'agit là d'une vaste problématique qui a été envisagée selon deux techniques différentes. L'approche classique est la plus utilisée et consiste à rassembler un corpus le plus représentatif possible de l'état de la langue qui veut être modélisé et à estimer les probabilités de chaque mot par maximum de vraisemblance (Agresti, 2002), c'est-à-dire par comptage des occurrences.

Historiquement, cette méthode a été utilisée par Thorndike, dont nous avons déjà décrit précédemment la liste (cf. 2.2.1), ou encore Kučera et Francis (1967), qui ont travaillé avec un corpus anglais d'un million de mots (le corpus *Brown*). Son défaut est qu'elle présente des limitations dont certaines étaient déjà envisagées par Thorndike lui-même. Ainsi, lorsque le nombre total de mots du corpus est insuffisant, l'estimation par maximum de vraisemblance produit des estimations erronées pour les mots peu fréquents (Zevin et Seidenberg, 2002, 4). De plus, le lexique, modélisé à partir de textes représentatifs d'une certaine langue écrite, fait habituellement fi de nombreux termes plus familiers, argotiques ou issus de la langue orale (Blair *et al.*, 2002). Enfin, Gougenheim *et al.* (1964) ont montré que les fré-

quences de certains termes, qu'ils désignent sous l'appellation « mots disponibles », étaient mal évaluées.

Ces différentes faiblesses ont depuis fait l'objet d'attentions, de sorte que les listes actuelles reposent sur des corpus nettement plus gros et qui comprennent des documents plus variés au niveau du registre de langue. Par ailleurs, avec le développement des Technologies de l'Information et de la Communication (TIC), certains auteurs (Blair *et al.*, 2002; New *et al.*, 2004) ont envisagé une autre façon de calculer les fréquences des mots de la langue, qui répondent aux faiblesses de l'approche classique. Il s'agit de remplacer les fréquences par le nombre de résultats rapportés par un moteur de recherche tel que Google pour un mot donné. Blair *et al.* (2002, 287) considèrent que le web constitue une ressource à la fois représentative, complète, contemporaine et facilement accessible, mais aussi que la quantité de résultats est un estimateur valide de la fréquence d'un terme dans la langue. Pour 400 mots anglais, ils ont obtenu un haut degré de corrélation entre cette nouvelle mesure et les fréquences de deux corpus classiques (Baayen *et al.*, 1993; Kučera et Francis, 1967). New *et al.* (2004) mettent toutefois en garde contre le fait que la présence d'un mot dans un document n'est pas équivalente à son nombre d'occurrences, mais aussi contre l'excès de représentation des sites à contenu sexuel. Enfin, notons que cette nouvelle vision ne permet pas d'obtenir des fréquences relatives pour chaque mot, puisque le nombre total de documents indexés est incertain.

Une seconde problématique liée à la conception de listes fréquentielles et à leur emploi en lisibilité est de se mettre d'accord sur l'unité linguistique dont il convient d'estimer la fréquence. En effet, la grande majorité des approches rassemblent sous une même forme graphique non seulement les différentes acceptations d'un même lexème, mais aussi les différents homographes. Nous avons déjà souligné l'intérêt qu'il y aurait en lisibilité à prendre en compte la fréquence des sens plutôt que celle des formes lexicales, même si cette démarche reste actuellement aux limites des techniques actuelles de TAL, en particulier pour le français.

À l'inverse, on peut s'interroger sur l'intérêt de travailler avec les formes fléchies en lisibilité, sachant que l'identification de la majorité des formes fléchies ne devrait pas poser de problèmes à un apprenant qui connaît déjà le lemme. Par ailleurs, qu'en est-il des mots composés ? Dans la majorité des formules de lisibilité, c'est la fréquence de chacun élément constitutif qui est pris en compte. Toutefois, une approche plus respectueuse du processus d'acquisition du vocabulaire en L2 devrait les considérer comme une unité lexicale unique (Clark, 1981).

Enfin, un dernier aspect, doublement problématique pour les formules de lisibilité, est la question des noms propres. D'une part, l'estimation des fréquences des noms propres est

complexe, sachant que ce « lexique » évolue sans cesse et qu'il faudrait pouvoir désambiguïser les homonymes. D'autre part, il n'est pas évident que l'effet de fréquence observé pour les noms communs s'applique de la même manière aux noms propres.

À ce propos, Morriss and Halverson (1938, cités par Chall, 1958, 25) distinguent déjà deux types de noms propres : ceux à contenus et les autres. Les premiers réfèrent à des personnes, lieux, événements et autres réalités historiques ou contemporains, tandis que les seconds sont co-textuels et servent à nommer des personnages fictionnels. Cette approche nous semble toutefois critiquable. Si un nom qui n'est pas référentiel par rapport au réel ne nécessite pas de connaissances préalables à l'univers du récit pour être appréhendé, une fois le récit commencé, il devient nécessaire d'associer une série d'événements et de traits à ce nom, sans quoi le texte risque d'être mal compris. Par ailleurs, il est aussi possible de faire référence à un personnage de récit depuis un autre texte et il nous semble qu'on retombe alors dans le cas référentiel.

Malgré ces bonnes raisons de tenir compte des noms propres, de nombreux auteurs en lisibilité considèrent qu'ils ne présentent pas de difficultés particulières à la lecture. Ainsi, Flesch (1948, 1950) les englobe dans ses deux catégories « intérêt humain » et « mots définis » supposés faciliter la lecture ; Gunning (1952, 38) considère les noms propres de trois syllabes ou plus comme simples, à la différence des noms communs de même longueur, et Daoust *et al.* (1996, 212) estiment eux aussi que les noms propres doivent être considérés comme connus. Il est toutefois intéressant de citer le cas de Fry qui, ne donnant initialement aucune consigne sur le sujet, s'est ensuite aligné en fonction de cette vision dominante, avant de se rétracter (Fry, 1977, 244). Il ne voit finalement pas pourquoi, pour un enfant, « Joe » ne serait pas plus facile à lire que « Joseph ». Mais surtout, il reconnaît que le désir de ne pas prendre en compte des noms propres venait des éditeurs qui avaient sinon plus de mal à respecter le niveau de lisibilité nécessaire pour vendre leurs manuels.

Les différentes mesures de la fréquence lexicale : : Si l'effet de fréquence a fait l'objet d'un certain consensus depuis les années 20, la manière dont on a paramétrisé le phénomène a considérablement évolué depuis les débuts de la lisibilité. À l'origine, les chercheurs devaient établir manuellement la liste des termes présents dans chaque texte de leur corpus. Par conséquent, calculer les probabilités précises de chaque mot aurait constitué une gageure supplémentaire et les premiers travaux se contentaient de vérifier le nombre d'absents d'une liste comprenant les X mots les plus fréquents de la langue. Par la suite, on a commencé à affiner cette façon de faire en rassemblant les mots de fréquences proches en tranche de X mots, avant que l'arrivée de l'ordinateur rende possible la prise en compte de la probabilité précise de chaque mot au sein des formules de lisibilité. Ces différentes étapes sont détaillées dans les deux sections suivantes, qui traitent respectivement des approches

par liste et de différentes manières d'utiliser les fréquences lexicales.

Le pourcentage d'absents d'une liste de référence : Un grand nombre de chercheurs (Dale et Chall, 1948a; Gray et Leary, 1935; Harris et Jacobson, 1974; Henry, 1975; Johnson, 1927; Lively et Pressey, 1923; Lorge, 1939; Vogel et Washburne, 1928; Washburne et Morphett, 1938) ont basé leur formule sur l'opposition entre mots « faciles » et mots « difficiles », la qualité de « mot facile » étant généralement liée à la présence de celui-ci dans une liste de mots considérés comme faciles, parce que les plus courants. De nombreuses listes ont été utilisées dans ce but, dont les plus connues sont celles de Thorndike, les lexiques de 769 mots et de 3 000 mots de Dale ou encore le *Français Fondamental* de Gougenheim *et al.* (1964)[3].

L'existence de multiples listes s'explique d'abord par le souci de pallier les faiblesses des réalisations précédentes ou d'adapter celles-ci à un autre contexte d'utilisation. Ainsi, Dolch (1948, 250) critique l'emploi de la liste de Thorndike pour estimer le vocabulaire connu des enfants, car il estime que les 1 000 premiers mots comprennent de nombreux termes davantage compréhensibles par des adultes. Une autre raison de cette multiplicité est l'existence d'au moins deux techniques pour construire ces listes. À côté de l'approche par corpus décrite ci-dessus, Dale a eu recours à une démarche qui paraît moins généralisable, mais plus adaptée à son objet. Ainsi, sur base d'un inventaire de 10 000 termes, il a testé le bagage lexical d'un échantillon d'élèves en quatrième année primaire et n'a retenu que les mots connus d'au moins 80% d'entre eux (Dale et Chall, 1948a). Enfin, certains chercheurs ont estimé nécessaire de créer de nouvelles listes afin de rester en phase avec l'état de la langue.

Le pourcentage d'absents d'une liste de référence a prouvé sa validité à maintes reprises comme mesure de la complexité lexicale. Ainsi, la formule de Dale et Chall (1948a), qui repose sur le nombre d'absents de la liste de 3 000 mots de Dale, a longtemps été considéré comme l'une des meilleures en éducation (Klare, 1988, 18). Klare (1984, 686), entre autres, pense que l'emploi d'une liste, si elle demande plus de travail que le comptage des syllabes, constitue un indice plus précis de la lisibilité.

À l'inverse, Elley (1969) considère que le nombre de mots absents d'une liste de référence est un indice grossier, car il s'agit d'une classification binaire sans degrés intermédiaires. Elle recommande de prendre en compte plus finement les différences d'emploi des mots dans la langue en se basant sur des comptages. Cette position tranchée semble pouvoir être remise en question en regard des résultats d'Henry (1975, 75), qui montre que les deux approches obtiennent en réalité des résultats relativement similaires.

3. Notre propos n'est pas ici de retracer un historique des différentes listes utilisées en lisibilité, d'autant que ces aspects ont déjà été largement couverts par le Chapitre 2. Nous renvoyons le lecteur intéressé par cette question à l'excellente synthèse de Chall et Dale (1995, 119-128).

Il faut toutefois reconnaître à cette variable une faiblesse importante relevée par Dale et Chall (1948a) ou Harris et Jacobson (1974) : son pouvoir de discrimination oscille en fonction du nombre de mots présents dans la liste. Moins il y en a, et plus les textes analysés comprendront de « mots difficiles ». Ainsi, Flesch (1943, cité par Dale et Chall, 1948a, 12) montre que les formules basées sur la courte liste de Dale (769 mots) éprouvent des difficultés à discriminer des textes plus complexes que le *grade* huit. Or, la difficulté moyenne des textes pour adultes se situe autour du 8e et du 9e *grade*.

Harris et Jacobson (1974, 1) parviennent à une conclusion similaire en concevant une formule basée sur le vocabulaire des *readers*, des livres simplifiés classés par niveau scolaire. Ils observent que la liste ne comprenant que les mots de la première année, c'est-à-dire 335 lemmes et leurs formes fléchies, n'est pas suffisamment discriminante pour les textes de niveaux plus élevés. Inversement, la liste rassemblant les mots des trois premières années (environ 6000 formes) peine à discriminer les textes faciles (mais les auteurs estiment qu'elle serait sans doute plus appropriée pour une formule travaillant sur des niveaux supérieurs aux primaires). Au final, ils recourent à une liste de 912 lemmes et 1880 formes fléchies attestées dans les *readers* de 1re et 2e année.

Pour conclure, notons que cette variable reste utilisée aujourd'hui, par Schwarm et Ostendorf (2005, 526) ou Kate *et al.* (2010), sous la forme d'un pourcentage de mots hors vocabulaire. Les premiers ont calculé ce pourcentage à la fois au niveau des occurrences et au niveau des « types ». Dans ce dernier cas, les mots « difficiles » répétés dans un même texte ne sont comptés qu'une fois. Cette façon de faire apparaît préférable, du moins d'après Nyman *et al.* (1961) et Harris et Jacobson (1974) qui obtiennent de meilleures performances sans duplication.

Statistiques des fréquences lexicales : Parallèlement au pourcentage d'absents d'une liste de fréquences lexicales, Lively et Pressey (1923) ont exploré une seconde voie pour évaluer la charge lexicale d'un texte. Elle consiste à se baser sur la fréquence de chaque mot afin d'obtenir une distribution d'échantillonnage fréquentielle pour le texte à analyser. Celle-ci n'étant pas aisément manipulable, on a généralement résumé l'ensemble des informations représentées par cette distribution sous la forme d'une statistique ou d'une autre.

Lively et Pressey (1923, 393) emploient la médiane des indices de fréquences de tous les « types » du texte, lesquels sont spécifiés dans la liste de Thorndike. Ils considèrent qu'il s'agit probablement là de la meilleure mesure de la charge lexicale. Pour leur part, Patty et Painter (1931) préfèrent effectuer la moyenne arithmétique des indices de Thorndike sur la base des *tokens* du texte. Ces deux statistiques – la moyenne et la médiane – seront les deux statistiques prédominantes dans les études de lisibilité qui suivent.

Le revers de cette paramétrisation est son aspect fastidieux, en particulier pour l'utilisateur des formules. Aussi, si les chercheurs de la première moitié du 20e siècle (par exemple, Lorge (1939)) testent cette variable, ils ne l'insèrent pas dans leur formule. D'autres (Bormuth, 1966; Forbes et Cottle, 1953; Tharp, 1939) vont simplifier le problème en discrétisant les fréquences. Le principe est de diviser les mots de la liste de référence, ordonnés selon leur fréquence, en n tranches de longueur égale. On obtient alors une distribution discrète, plus facilement manipulable, et dont il est plus simple de calculer la moyenne ou la médiane. Tharp (1939, 171) critique cette façon de faire qui revient à postuler que, dans le cas de classes de 1 000 mots, le 1 000e mot n'est pas plus complexe que le vocable le plus fréquent de la langue.

Avec l'arrivée de l'ordinateur, il est désormais possible de s'affranchir de ces ennuyeux calculs et d'estimer directement diverses statistiques de tendance centrale ou de dispersion à partir des fréquences des mots du texte à analyser. Étonnamment, cette approche n'a cependant guère été utilisée (Crossley *et al.*, 2007, 2008; Graesser *et al.*, 2004; Stenner, 1996; Weir et Ritchie, 2006). Notons que McCallum et Peterson (1982) et, par la suite, Stenner (1996) privilégient le logarithme des fréquences aux fréquences brutes. Cette transformation permet de réduire l'asymétrie de la distribution des fréquences et correspondrait davantage à la relation qui existe entre les fréquences et les temps de latence observés dans les diverses tâches de reconnaissance des mots (Balota et Chumbley, 1984; Just et Carpenter, 1980; Whaley, 1978).

Variantes de l'effet de fréquence Avant de clore la description de cette première famille de prédicteurs basée sur l'effet de fréquence, il est intéressant d'envisager rapidement trois autres variantes de ce facteur. La première est due à Elley (1969) qui propose de ne prendre en compte que la fréquence des noms. En effet, certaines classes de mots seraient représentées de la même façon dans les textes faciles et difficiles et il ne servirait à rien de les traiter. De plus, les résultats à des tests de closure indiquent que ce sont les noms qui sont les plus difficiles à compléter correctement, ce qui signifie qu'ils forment la partie la moins redondante d'une phrase et donc, la partie la plus cruciale pour en saisir le sens. Son hypothèse semble corroborée par un taux de corrélation avec le critère supérieur à celui de facteurs classiques (Elley, 1969, 419).

Une deuxième variante assez proche de celle d'Elley est attribuée à Bormuth (1969a, 17), lequel calcule un indice de fréquence à partir des mots lexicaux (c'est-à-dire les noms, les verbes, les adverbes et les adjectifs). Cet indice correspond à la somme des indices de fréquence de la liste de Thorndike divisée par le nombre de mots lexicaux. On trouve un argument en faveur de cette variable au milieu des critiques de Selzer (1981, 29) à l'encontre de la lisibilité. Celui-ci indique en effet que les mots grammaticaux ont tendance à occulter,

au sein d'une mesure de fréquence, la complexité des termes lexicaux où se situe généralement la difficulté. Autrement dit, un texte avec de nombreux mots grammaticaux et quelques termes rares ne sera pas détecté comme plus difficile qu'un texte avec des noms et verbes relativement simples. C'est pourquoi une variable de la fréquence lexicale des mots lexicaux semble justifiée. Notons que Bormuth (1969a) définit aussi un indice similaire pour les mots grammaticaux.

Enfin, Chall et Dale (1995, 118) citent une troisième idée qui nous paraît fort intéressante, mais n'a jamais été testée, à notre connaissance. Il s'agirait de repérer les mots jouant un rôle majeur dans un texte, puisque leur connaissance est primordiale pour appréhender les idées principales véhiculées. Dès lors, seule la fréquence de ces termes clefs serait prise en compte au sein des mesures de la complexité lexicale. On pourrait aussi imaginer une approche plus nuancée, où chaque mot se verrait attribuer un poids en fonction de son importance dans le texte. Le problème est ici d'arriver à distinguer les mots jouant un rôle majeur des autres.

Conclusion : Au terme de cette longue analyse des variables basées sur la notion de fréquence, on s'aperçoit qu'il s'agit de variables efficaces et largement employées, aussi bien via le nombre d'absents d'une liste de référence que de statistiques de fréquence. Nous avons dès lors décidé de retenir plusieurs d'entre elles à fins d'expérimentations, malgré que Laroche (1979) doute de la pertinence d'appliquer cette famille de mesures au contexte L2. En effet, d'une part nous ne disposons pas d'assez de textes pour estimer fiablement les fréquences lexicales propres à l'environnement FLE ; d'autre part, nous avons expliqué précédemment pourquoi les fréquences en L1 pourraient être de meilleures estimations des fréquences L2 que ce que pense Laroche (1979). Nous envisageons toutefois deux autres approches fréquentielles dans la suite de cette section : la vraisemblance textuelle et la familiarité, où est justement discutée plus en détail la question de la fréquence objective.

Mesures de la vraisemblance textuelle

Avec le développement de la lisibilité computationnelle est apparue une nouvelle famille de prédicteurs reposant également sur l'effet de fréquence, mais qui aborde celui-ci à l'aide d'un modèle issu du TAL : les modèles de langue, aussi appelés modèles n-grammes (*cf.* Jurafsky et Martin 2009). Ceux-ci reposent aussi sur des fréquences calculées par maximum de vraisemblance, mais ne les utilisent pas dans le même but. Si les diverses variables que nous avons décrites précédemment constituent des indicateurs d'une tendance de la distribution fréquentielle d'un texte, appliquer un modèle de langue à un texte revient à en éprouver sa vraisemblance, c'est-à-dire à se poser la question suivante : quelle est la probabilité qu'une telle suite de mots existe dans la langue ?

Les modèles n-grammes en lisibilité : Formellement, la probabilité conjointe d'une suite de n mots peut se calculer comme le produit des probabilités conditionnelles de chaque mot, selon la règle des probabilités composées :

$$P(w_1, w_2, \cdots, w_n) = P(w_1)P(w_2 \mid w_1) \cdots P(w_n \mid w_1, w_2, \cdots, w_{n-1}) \tag{3.1}$$

Toutefois, il est bien connu que les probabilités conditionnelles situées en fin d'expression deviennent rapidement inestimables, car elles correspondent à des suites de mots trop rarement observées. Dès lors, la probabilité d'une suite de mots peut être approximée à l'aide d'un modèle n-gramme, où seuls les $n-1$ mots précédents sont pris en considération dans chaque probabilité conditionnelle. Cette approximation de l'Équation 3.1 s'exprime ainsi, dans le cas d'un modèle d'ordre 3, aussi appelé modèle trigramme :

$$P(w_1, w_2, \cdots, w_n) = P(w_1)P(w_2 \mid w_1) \prod_{i=3}^{n} P(w_i \mid w_{i-1}, w_{i-2}) \tag{3.2}$$

En lisibilité, la nature de cette variable varie en fonction de l'ordre du modèle. Dans le cas le plus simple (où $n = 1$), qu'on désigne sous le terme d'unigramme, ce prédicteur se situe uniquement à un niveau lexical. La principale différence avec, par exemple, la moyenne des fréquences relatives décrite précédemment, tient à ce que le modèle unigramme n'additionne pas les probabilités de chaque mot, mais les multiplie. L'effet d'un seul mot rare n'est pas noyé au sein de la masse des autres observations, mais influence considérablement le résultat final. Pour les modèles d'ordre supérieur, il faut être attentif au fait que cette variable saisit alors un mélange de régularités lexicales et syntaxiques de la langue.

La première apparition des modèles de langues en lisibilité est due à Si et Callan (2001), dont nous avons déjà présenté les travaux précédemment (*cf.* Section 2.2.8). Comme nous l'avons expliqué, ces deux auteurs n'utilisent toutefois pas un modèle de langue unique , mais entraînent un unigramme pour chaque niveau de difficulté. Cette approche, qui nous semble particulièrement intéressante pour le FLE, a été reproduite et améliorée par Collins-Thompson et Callan (2004a, 2005), ainsi que par Heilman *et al.* (2007) (*cf.* Section 2.2.8).

Une approche plus simple consiste à se limiter à un seul modèle unigramme pour l'ensemble des niveaux. Elle a été implémentée par Pitler et Nenkova (2008, 189), qui obtiennent une corrélation significative pour ce prédicteur ($r = 0,37$ sur un corpus et $r = 0,45$ sur l'autre). Il faut cependant souligner qu'elles ne normalisent pas le résultat obtenu en fonction de la longueur des textes. De même, nous avons précédemment (François, 2009a,b) considéré cet indice comme notre meilleur ($r = -0,59$).

Les expérimentations sur des modèles d'ordre supérieur sont dues à Kate *et al.* (2010). Ils

proposent de plus une manière simple, mais efficace de normaliser la probabilité obtenue en fonction de la taille du document D (Kate *et al.*, 2010, 549) :

$$NormP(D) = (P(D \mid M))^{\frac{1}{|D|}} \tag{3.3}$$

où M représente le modèle de langue utilisé, $\mid D \mid$ correspond au nombre de mots dans le document D et $P(D \mid M)$ est la probabilité obtenue pour ce document.

Mesures dérivées de modèles n-grammes : Certains des chercheurs qui font usage des modèles de langue en lisibilité ne se basent pas directement sur la probabilité fournie par l'Équation 3.2. Ils lui préfèrent d'autres mesures, telles que la perplexité ou un ratio de vraisemblance. Ces deux mesures ont été initialement explorées dans ce contexte par Schwarm et Ostendorf (2005) et reprises par plusieurs auteurs (Feng *et al.*, 2010; Kate *et al.*, 2010; Petersen et Ostendorf, 2006b). Schwarm et Ostendorf (2005) ont d'abord testé le ratio de vraisemblance RV qui vise à déterminer à comparer la probabilité qu'un document D appartienne à un niveau de difficulté c par rapport à la somme des probabilités d'appartenance aux autres niveaux. Il est défini formellement comme suit :

$$RV = \frac{P(D \mid c)P(c)}{\sum_{c_i \neq c} P(D \mid c_i)P(c_i)} \tag{3.4}$$

Suite à une série de tests, les auteurs ont préféré estimer la difficulté de textes en détournant la mesure de perplexité (PP), initialement conçue comme un moyen d'évaluation intrinsèque d'un modèle de langue. Ils appliquent à la lisibilité la définition classique de la perplexité (Schwarm et Ostendorf, 2005, 526) :

$$PP = 2^{H(D|c)} \tag{3.5}$$

où $H(D \mid c)$ est l'entropie d'une suite n d'observations relative à la classe c. Elle se calcule d'après la formule suivante :

$$H(D \mid c) = -\frac{1}{n} \log_2 P(D \mid c) \tag{3.6}$$

Dans ce cas, c représente le modèle de langue entraîné sur des textes d'un niveau particulier, mais pourrait tout aussi bien utiliser un modèle de langue unique pour l'ensemble des textes du corpus.

Notons que Kate *et al.* (2010), qui ont aussi intégré la perplexité comme variable au sein d'une formule de lisibilité, concluent qu'il s'agit de leur meilleur prédicteur. Ils apportent toutefois quelques modifications à la mesure, travaillant avec des modèles M entraînés par

genre et non par niveau de difficulté et utilisant le ratio suivant pour comparer la perplexité d'un genre donné par rapport aux autres (Kate *et al.*, 2010, 550) :

$$R(M_i \mid D) = \frac{L(D \mid M_i)}{\sum_{i=1}^{G} L(D \mid M_i)} \tag{3.7}$$

où $L(D \mid M_i)$ correspond à la perplexité d'un document étant donné le modèle M associé au genre i, sachant qu'il y a G genres et donc, autant de modèles. Notons que cette formule postule que les probabilités a priori des G modèles de langues sont équiprobables.

En conclusion de cette discussion sur l'usage des modèles de langues dans les formules de lisibilité, notons qu'à notre connaissance, personne n'a encore comparé précisément l'ensemble de ces différentes mesures entre elles, ni avec les techniques plus classiques issues de la statistique descriptive. Par contre, plusieurs comparaisons locales ont été effectuées, parmi lesquels l'une des plus intéressantes est rapportée par Collins-Thompson et Callan (2004a) qui concluent à la supériorité de leur approche par unigrammes sur la moyenne du logarithme des fréquences (Stenner, 1996). Ils ajoutent que les modèles de langues sont particulièrement plus performants sur des textes très courts, lesquels posent des problèmes à la majorité des variables en lisibilité.

Mesures de la familiarité lexicale

Comme nous l'avons dit précédemment, les variables basées sur la fréquence des mots dans la langue écrite, autrement dit sur la fréquence objective, ont connu un large succès en lisibilité. Cependant, certains chercheurs pensent que le recours à des listes lexicales ne représente qu'une approximation grossière de l'impact de la fréquence d'occurrence des mots sur les performances de lecture d'un individu donné. En réalité, il serait plus adéquat de tenir compte de la familiarité des mots, c'est-à-dire de la fréquence subjective. Celle-ci serait liée, pour Gernsbacher (1984), au nombre de fois qu'une personne a lu, entendu, produit ou écrit un mot donné.

Diverses études ont confirmé cet effet de la fréquence subjective. Solomon et Postman (1952) conçoivent une expérience où dix mots étrangers sont lus un nombre de fois variable par des sujets avant que leur temps de reconnaissance minimal ne soit mesuré. Les résultats indiquent clairement que la fréquence d'exposition d'un sujet donné à un mot est inversement corrélée au temps minimal nécessaire pour le reconnaître. Ils apparaissent d'autant plus intéressants pour notre étude, puisqu'ils s'appliquent à des mots d'une L2. Par la suite, Gernsbacher (1984) et Connine *et al.* (1990) effectuent une série d'expérimentations qui montrent que (1) les effets de fréquence et de familiarité co-existent ; (2) si la familiarité des mots fréquents reste relativement stable d'un individu à l'autre, justifiant l'approximation

par la fréquence objective, pour les mots de basse fréquence, elle varie fort d'une personne à l'autre (3) en conséquence, la familiarité serait un meilleur prédicteur du temps de réponse à diverses tâches classiques que la fréquence dans le cas des mots peu communs.

Sur la base de l'effet de familiarité, nous pouvons postuler hypothèse suivante :

Hypothèse 3.2 *Plus un individu a rencontré un mot donné, plus la lecture de celui-ci est facilitée. Par conséquent, plus il y a de mots familiers dans un texte, plus la lecture de celui-ci est aisée.*

Reste à évaluer cette familiarité ! Gernsbacher (1984), par exemple, s'y prend en demandant à une série de sujets de juger de leur familiarité avec un ensemble de mots sur une échelle à 7 niveaux. Au-delà des critiques intrinsèques concernant la fiabilité de cette démarche, il est assez évident que celle-ci pose des problèmes supplémentaires en lisibilité, où l'objectif est de modéliser un lecteur prototypique d'une population et non de s'arrêter aux détails individuels. On pourrait évidemment recourir à une enquête, similaire à celle de Gernsbacher (1984), sur des sujets appartenant à la population d'intérêt et prendre la moyenne de leurs réponses pour chaque mot. Toutefois, cela demanderait beaucoup d'efforts pour un résultat probablement assez proche des fréquences objectives estimées sur corpus.

Si cette dernière remarque est vraie pour la lisibilité d'une L1, l'effet de familiarité nous apparaît plus pertinent dans le cadre de la lecture en L2, comme l'avait déjà souligné (Laroche, 1979), en particulier lorsque l'acquisition de cette L2 s'opère essentiellement dans un contexte de classe. En effet, dans ce cas, le degré d'exposition des apprenants au lexique procède majoritairement des matériaux pédagogiques qui leur sont soumis. Dès lors, il est possible d'évaluer approximativement la familiarité de ces apprenants au lexique de la L2 au départ d'un corpus de documents destinés à l'enseignement de cette L2, pour autant que l'on dispose d'un nombre suffisant de textes. Autrement dit, on se baserait sur les fréquences lexicales, estimées non pas sur un corpus de L1, mais sur un corpus de L2.

Il faut distinguer, d'une part, cette approche du modèle unigramme de Collins-Thompson et Callan (2004a), qui estiment, pour chaque mot, sa distribution sur les niveaux de difficulté, mais pas spécialement en se limitant à des textes de L2. D'autre part, la familiarité ne doit pas être confondue avec l'âge d'acquisition, qui est décrit ci-après. Notons enfin que la fréquence subjective peut aussi être intégrée dans une formule de lisibilité adaptative, où le vocabulaire des textes lus par un apprenant précis serait mémorisé par un système. Les prédicteurs lexicaux de cette formule pourraient alors être adaptés en fonction de la fréquence d'occurrence des mots rencontrés jusque-là par l'apprenant. Nous n'avons toutefois pas utilisé ce type de mesure pour notre étude, à cause de la taille trop réduite de notre corpus.

Mesures de la diversité lexicale

Cette quatrième famille de prédicteurs lexicaux s'appuie non plus sur la fréquence des mots dans la langue, mais sur leur fréquence au sein du texte à lire. Autrement dit, ce type de variable vise à prendre en compte **l'effet de répétition des mots** au sein d'un texte et repose sur l'hypothèse suivante :

Hypothèse 3.3 *La lecture d'un mot cible est facilitée si ce même mot a été rencontré auparavant. Par conséquent, plus il y a de mots qui se répètent dans un texte, plus la lecture de celui-ci est aisée.*

L'effet de répétition (en anglais, *repetition priming effect*) est, d'après Ferrand (2007, 160), « l'un des plus robustes effets observés en lecture ». Dans une synthèse récente sur le sujet, Bowers (2000) rappelle qu'il faut distinguer deux types d'effet de répétition. Le premier agit sur le long terme et peut perdurer jusqu'à quelques jours, tandis que le second durerait moins de 500 ms, d'après Ferrand (1996). En conséquence, c'est surtout le premier qui mérite d'être pris en compte en lisibilité, étant donné qu'un laps de temps de 500 ms ne permet pas d'appréhender plus de quelques mots.

L'effet de répétition à long terme a fait l'objet de nombreuses études qui ont révélé trois caractéristiques notables. Tout d'abord, il serait plutôt de nature orthographique, même s'il existe une hypothèse alternative, selon laquelle il s'expliquerait par la récupération en mémoire d'une trace épisodique (Tenpenny, 1995). Selon l'hypothèse de l'activation orthographique, défendue entre autres par Bowers (2000), l'effet de répétition serait dû à l'activation récente de l'entrée lexicale concernée dans le lexique mental, ce qui rendrait son accès plus aisé durant une certaine période. Dans la vision basée sur la mémoire épisodique, la reconnaissance des mots s'effectuerait toujours via une expérience préalable précise et l'effet de répétition viendrait du fait qu'un épisode plus récent est plus accessible que ceux qui sont plus anciens (Tenpenny, 1995, 340).

Un second aspect est que l'effet de répétition interagit avec l'effet de fréquence, ce qui signifie concrètement qu'il est plus important pour les mots rares que pour les mots fréquents (Forster et Davis, 1984; Scarborough *et al.*, 1977). Cette caractéristique nous importe davantage que la première, puisqu'elle révèle que l'effet de répétition joue un rôle plus important pour les mots rares. Or, il s'agit justement des termes lexicaux les plus susceptibles de poser problème à des lecteurs dans une langue étrangère. Il nous semblerait dès lors opportun d'expérimenter des mesures de la diversité lexicale qui se basent sur ces termes infréquents. Cela explique peut-être aussi pourquoi une mesure classique de la diversité lexicale telle que le ratio entre les « types » et les tokens (*type token ratio*) s'est rarement révélé un prédicteur

efficace en lisibilité.

Enfin, la troisième caractéristique intéressante de l'effet de répétition concerne les non-mots. Un non-mot est une chaîne de caractères qui ne respecte pas les règles phonologiques d'une langue. Il convient de les distinguer d'un pseudo-mot, une chaîne de caractères respectant ces règles, mais qui ne porte aucun sens. Scarborough *et al.* (1977), entre autres, a obtenu un effet de répétition significatif aussi pour les non-mots lors d'une tâche de décision lexicale. Dans le cadre de la lecture en L2, cela signifie que les mots encore inconnus du lecteur devraient bénéficier de cet effet, quand bien même le système phonologique de la langue cible n'aurait pas été complètement assimilé.

Il faut cependant noter que cet effet facilitateur intervient au niveau de la reconnaissance des mots. Cela ne signifie pas qu'un mot inconnu du lecteur sera pour autant mieux compris s'il a été rencontré auparavant dans le texte. Se pose aussi la problématique de la polysémie : qu'en est-il lorsque deux homographes sont présents dans le même texte avec un sens différent ? Il semblerait que l'ensemble des sens soient activés par la forme graphique et que celui qui a le plus haut taux d'activation, en fonction de sa fréquence a priori ainsi que du contexte, soit privilégié (Just et Carpenter, 1980; Nievas et Mari-Beffa, 2002). Par ailleurs, des études (Meyer et Schvaneveldt, 1971; Neely, 1991) ont montré que l'effet de répétition existait également pour un mot sémantiquement relié à la cible (en anglais, *semantic priming*). On peut donc supposer que, dans le cas d'homographes, l'effet de répétition jouera, mais dans une moindre mesure.

Au-delà de la reconnaissance des mots, certains chercheurs rapportent un effet empirique de la répétition des mots pour des processus de niveaux supérieurs. Ainsi, Richaudeau (1984, 138) note que la répétition d'un mot dans une phrase améliore le score de mémorisation de celle-ci. Toutefois, de manière générale, les mesures de la redondance textuelle ont été assez controversées en lisibilité, comme nous allons le voir dans la section suivante, qui traite de la plus employée d'entre-elles : le TTR.

Le Type Token Ratio Utilisée dès 1923 par Lively et Pressey (1923), cette mesure de la diversité lexicale correspond au pourcentage de mots différents dans un texte. Il semblerait que dès ses débuts, elle n'ait pas fait l'unanimité. Si c'est avec le TTR que Vogel et Washburne (1928) obtiennent la meilleure corrélation avec leur critère, il n'en va pas de même pour Lively et Pressey (1923, 392). Ceux-ci observent que si cet indice permet bien de distinguer les différences en termes de niveau de difficulté au sein de textes littéraires, la diversité lexicale s'applique mal aux textes scientifiques. Aussi complexes que les textes littéraires les plus ardus du corpus, les textes scientifiques obtiennent pourtant un résultat nettement inférieur en termes de diversité lexicale, parce qu'ils réemploient sans cesse les mêmes termes techniques, à côté de mots relativement simples. Aussi, le TTR ne serait pas adéquat pour tous

les types de textes, mais plutôt adapté aux textes littéraires et, en particulier, aux matériaux destinés aux enfants (Chall, 1958, 41).

Le TTR est repris par Dolch (1928, 172), qui montre déjà une limitation fondamentale de cette mesure : sa normalisation. Cette problématique est très bien décrite quelques années plus tard par Patty et Painter (1931, 129), qui déconseillent l'emploi d'un extrait pour estimer la diversité lexicale d'un ouvrage :

> *The different words in a one-thousand-word unit from a short text bear a lower ratio to the total one thousand than the different words in the entire text might bear to the total words in it. In other words, the longer the text the greater is the probability of its having a high percentage of word repetition.*

Derrière cette critique, il faut distinguer deux aspects différents. La remarque principale de Patty et Painter (1931) porte avant sur un problème d'estimation statistique. Ils notent que comme le TTR dépend de la longueur du texte à analyser, recourir à un échantillon de 1 000 mots pour calculer la difficulté d'un livre ne fournit pas une estimation correcte du TTR pour cet ouvrage.

Toutefois, maintenant que les techniques informatisées autorisent l'analyse complète de n'importe quel livre, cette problématique n'est plus actuelle et c'est le second aspect de la critique de Patty et Painter (1931) qui s'impose, c'est-à-dire la dépendance du TTR envers la longueur du texte à analyser. À notre connaissance, cette question n'a toujours pas été réglée élégamment, puisque le manuel d'un logiciel de statistique linguistique tel que WordsmithTool nous met en garde [4] :

> *Le TTR varie énormément en fonction de la longueur du texte (ou du corpus de texte) étudié. Un article de 1 000 mots peut avoir un TTR de 40%, un texte plus court avoir un ratio de 70% et un texte de 4 millions de mots, quant à lui, peut avoir un ratio de 2%. [...] si vos recherches s'orientent sur le texte par opposition à la langue, vous serez amené à traiter des textes de longueurs différentes, et le TTR conventionnel ne vous sera pas d'une grande aide.*

Malgré ce défaut conséquent, le TTR a continué à être utilisé par plusieurs chercheurs, dont Washburne et Morphett (1938), Henry (1975) ou Kemper *et al.* (1993). Graesser *et al.* (2004) en emploient une variante située dans le même ordre d'idées que l'approche des fréquences lexicales par Elley (1969). En effet, ils ne calculent le TTR que pour les mots lexicaux et pas les mots fonctionnels. De plus, ils définissent aussi un TTR spécifique aux noms, ainsi

4. Citation extraite du manuel en ligne de WordSmithTool et consultable en ligne à l'adresse `http://www.lexically.net/downloads/version4/html/francais/index.html?type_token_ratio_proc.htm`

que son complément : l'ensemble des mots lexicaux excepté les noms.

Autres mesures de la diversité lexicale : À côté du TTR, d'autres mesures de la diversité ont été envisagées. Nous en citons rapidement deux. Patty et Painter (1931) proposent de normaliser la moyenne des indices de fréquence de la liste de Thorndike par ce qu'ils appellent le « range » des mots, à savoir le nombre de mots différents dans un échantillon de 1000 mots. L'intérêt de cette variable est qu'elle permet de distinguer deux textes dont la moyenne des fréquences est similaire, mais pas le TTR. Or, il est légitime de penser que la charge lexicale de celui qui comporte le plus de mots différents est supérieure.

La seconde approche cible cette fois la redondance textuelle. Carterette et Jones (1963) évaluent le taux de contrainte séquentielle au niveau des lettres sur la base de la théorie de l'information de Shannon et Weaver (1949). Ils proposent une formule dont les résultats, calculés sur des livres de 1re, 2e, 3e et 5e primaires, sont remarquablement corrélés avec la difficulté des textes : plus celle-ci augmente, plus la redondance diminue. Cette variable a été employée par Bormuth (1966, 100), sans qu'elle se révèle particulièrement informative ($-0,377 \leq r \leq 0,35$). C'est pourquoi nous ne l'avons pas reprise pour notre propre étude, préférant nous limiter à plusieurs versions du TTR, dont certaines sont normalisées d'après une technique présentée par après (*cf.* Section 3.3.1).

L'âge d'acquisition

Une autre variable cognitive importante pour expliquer diverses observations expérimentales en reconnaissance des mots est l'âge d'acquisition, c'est-à-dire l'âge auquel un mot a été appris par un individu donné. On trouve dans la littérature psycholinguistique des preuves expérimentales qui confirment l'existence d'un effet facilitateur lié à l'âge d'acquisition des mots, même si celui-ci a longtemps été peu considéré. En effet, l'âge d'acquisition étant fortement corrélé avec la fréquence objective, la plupart des études sur l'effet de fréquence n'ont pas contrôlé l'âge d'acquisition des mots testés, considérant que celui-ci était assimilable à la fréquence. Pourtant, Morrison et Ellis (1995) publient un article polémique qui avance que, dans des tâches de prononciation, une fois l'âge d'acquisition contrôlé, l'effet de fréquence disparaît, tandis que l'inverse n'est pas vrai. Par contre, en ce qui concerne la tâche de décision lexicale, qui s'apparente le plus à une lecture silencieuse puisqu'elle ne nécessite pas le recours au processus de vocalisation, ils observent des effets indépendants pour la fréquence et l'âge d'acquisition.

Les études qui suivent confirment cette coexistence des deux effets, aussi bien en prononciation (Brysbaert, 1996; Brysbaert *et al.*, 2000; Gerhand et Barry, 1998) que pour les tâches de décision lexicales (Bonin *et al.*, 2001; Brysbaert *et al.*, 2000; Gerhand et Barry, 1999). Dans

ce dernier cas, il s'agirait plus précisément d'une interaction : l'âge d'acquisition, comme la fréquence subjective, joue davantage pour les mots peu fréquents. Rappelons aussi que l'effet de la fréquence agirait plutôt lors de la reconnaissance de mots, tandis que l'âge d'acquisition affecterait davantage les tâches phonologiques. L'hypothèse que nous pouvons dès lors émettre est la suivante :

Hypothèse 3.4 *L'âge d'acquisition d'un mot affecte les performances du lecteur, même quand celui-ci est devenu adulte. Dès lors, plus les mots d'un texte ont été acquis tardivement par un individu, plus celui-ci sera susceptible d'éprouver des difficultés à la lecture dudit texte.*

À nouveau, comme il s'agit d'une variable propre à chaque lecteur, il n'est possible d'envisager l'âge d'acquisition en lisibilité qu'en référence à une population de référence. Dans ce cas, on peut soit recourir à une moyenne, soit à une distribution (c'est-à-dire une moyenne accompagnée de son écart-type) pour chaque élément du lexique. Reste encore à calculer ce(s) paramètre(s) en fonction de la population visée. À notre connaissance, seules deux études ont utilisé l'âge d'acquisition comme variable en lisibilité et en ont proposé une technique de paramétrisation. Graesser *et al.* (2004, 106) s'appuient sur la liste de 1 944 mots de Gilhooly et Logie (1980, cités par Graesser *et al.*, 2004), qui situe chaque mot sur une échelle de 1 à 7 [5]. Graesser *et al.* (2004) multiplient cette valeur par 100, mais leur article ne précise ni comment évaluer l'âge d'acquisition pour un texte, ni comment les mots non repris dans la liste sont traités.

La seconde technique, plus complète et complexe, est due à Kidwell *et al.* (2009). Ces chercheurs proposent de se baser sur un corpus de textes pédagogiques annotés en fonction de l'année scolaire où ils sont utilisés afin d'estimer automatiquement les distributions pour chaque mot du lexique. L'avantage de leur méthode est qu'elle permet de disposer d'une couverture lexicale nettement supérieure à celle qu'offre une liste. Par contre, elle ne modélise que l'âge d'acquisition des mots à l'écrit. Cela pose un problème pour les premières années de l'enfance durant lesquelles l'apprentissage du lexique passe uniquement par l'oral, puisque les premiers textes considérés remontent à la première primaire. En découlent des âges d'acquisition plus élevés en moyenne que ceux de la liste de Gilhooly et Logie, en particulier pour les mots qui sont appris entre deux et six ans. Néanmoins, Kidwell *et al.* (2009, 904) montrent que leur algorithme ne présente qu'une erreur moyenne absolue [6] de $0,5$ par rapport à la norme de Bristol (Stadthagen-Gonzalez et Davis, 2006).

5. D'après Ferrand (2007, 113), ce genre de liste est obtenue au moyen de mesures subjectives. Concrètement, cela consiste à demander à des adultes d'estimer l'âge auquel ils ont appris tel et tel mot. Malgré l'aspect subjectif de la méthode, celle-ci aurait une fiabilité très élevée.

6. L'erreur moyenne absolue permet de mesurer la moyenne des erreurs absolues commises par une série

Une fois les paramètres des V distributions normales obtenus (où V est le nombre de termes du lexique), les auteurs les utilisent pour prédire le niveau scolaire de nouveaux textes à l'aide d'un modèle de Rasche (1961). Celui-ci obtient de bons résultats, étant capable de prédire le niveau de nouveaux textes avec une erreur moyenne absolue de $1,40$. Ils montrent également que leur modèle est légèrement plus performant lorsqu'il emploie des distributions estimées sur corpus que lorsqu'il se base sur celles de Gilhooly et Logie ou celles de Stadthagen-Gonzalez et Davis (2006). Cette étude apparaît essentielle en ce qui concerne l'âge d'acquisition en lisibilité, puisqu'il s'agit non seulement de la première expérience en la matière, mais aussi parce qu'elle révèle que cette variable constitue un excellent prédicteur de la lisibilité en L1. Il est nettement corrélé avec le niveau des textes ($r = 0,59$), se révélant plus efficace que la moyenne du logarithme des fréquences ($r = 0,49$) et légèrement moins que le modèle unigramme de Collins-Thompson et Callan (2004a) ($r = 0,63$).

La situation est toutefois moins claire en ce qui concerne l'acquisition lexicale en L2. Dans ce cas, un réseau conceptuel a déjà été développé dans la L1, et on pourrait penser que ce sont surtout les nouvelles formes orthographiques associées à ces concepts qui font l'objet de l'apprentissage. Que doit-on considérer comme l'âge d'acquisition dans ce cas ? S'agit-il du moment où le concept est acquis dans la L1 ou bien du moment où le vocable de la L2 est appris ? Tout dépend de la façon dont les représentations conceptuelles et orthographiques des deux langues sont organisées.

Si l'on en croit Ellis et Ralph (2000), l'avantage des mots acquis plus tôt en L1 s'expliquerait par une plus grande plasticité du réseau neuronal dans les premières années. Peu à peu, celui-ci se stabiliserait et aurait plus de difficulté à assimiler un nouveau pattern. Il serait nécessaire que ce pattern soit rencontré un grand nombre de fois pour être intégré de manière équivalente aux termes acquis avant la stabilisation du réseau. Cette hypothèse peut être mise en relation avec celles qui défendent l'existence d'un âge critique pour l'acquisition optimale d'une L2 (Johnson et Newport, 1989; Snow et Hoefnagel-Höhle, 1978). D'après les travaux de Silverberg et Samuel (2004), il semblerait que l'âge d'acquisition joue différemment avant et après ce seuil critique, situé autour de la septième année.

Ces auteurs rapportent les résultats d'expériences bilingues basées sur la technique du « semantic priming ». Comme nous l'avons expliqué précédemment, il s'agit d'une tâche de décision lexicale dans laquelle le mot cible est précédé d'une amorce. Lorsque celle-ci est reliée sémantiquement à la cible, elle accélère la décision à propos de la cible. Dans le cas

de prédictions f_i en référence aux vraies valeurs y_i. Elle se calcule de la façon suivante :

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} \mid f_i - y_i \mid \tag{3.8}$$

bilingue, l'amorce est issu de la L2 et la cible est un terme de la langue maternelle des sujets.

Certaines études, parmi lesquelles Schwanenflugel et Rey (1986), ont validé la présence d'un effet facilitateur lorsque l'amorce est liée sémantiquement à la cible, ce qui semble indiquer l'existence d'un réseau conceptuel unique pour les deux langues. Cependant, Silverberg et Samuel (2004) avancent un tableau plus complexe. Avant le seuil critique, les représentations conceptuelles des mots dans les deux langues seraient partagées chez des bilingues, tandis que les représentations lexicales ne le seraient pas. Par contre, chez des bilingues plus tardifs, bien que tout aussi compétents, on ne peut conclure au partage des représentations sémantiques, alors que les représentations lexicales le seraient, car une forme L2 similaire à celle de la cible L1 tend à inhiber la reconnaissance de celle-ci.

Il nous semble que l'on peut conclure de ces faits que (1) l'âge d'acquisition du lexique d'une L2 doit plutôt être compris comme le moment où les mots de la L2, et non les concepts de la L1, sont acquis ; (2) dès lors, l'effet facilitateur n'apparaît que lorsque l'apprentissage de cette L2 a eu lieu avant le seuil critique. Par conséquent, cette variable ne semble pas adaptée à une formule de lisibilité spécialisée pour la L2, du moins en théorie. C'est pourquoi nous ne l'avons pas retenu pour nos expérimentations.

Les effets du voisinage orthographique

Cet autre effet lexical est très étudié depuis l'étude de Coltheart *et al.* (1977), qui ont défini N, une mesure de densité du voisinage orthographique correspondant à l'ensemble des mots de même longueur ne se différenciant que par une seule lettre (ex. pour SAGE : MAGE, SALE, etc.). À partir de cette mesure, ils espéraient observer un effet du nombre de voisins d'un mot sur la lecture de celui-ci. Cependant, ils ont obtenu un temps de décision plus long lors d'une tâche de décision lexicale pour des non-mots, et aucun effet pour les mots.

Ce résultat a ensuite été remis en question et, d'après Ferrand (2007, 129), les effets de voisinage sont encore largement débattus. En effet, ils conduisent à des résultats divers en fonction d'au moins trois variables : la langue considérée, la tâche demandée au sujet et le fait que l'on considère la densité du voisinage ou la fréquence des voisins. Andrews (1997) a proposé une synthèse de ces résultats divergents dans laquelle il montre qu'il est possible de réconcilier la plupart des études. Il révèle une opposition nette entre les observations concernant l'anglais, d'une part, et le français ou l'espagnol, d'autre part. En anglais, on note un effet de facilitation pour les termes ayant davantage de voisins orthographiques, alors qu'aucun effet similaire n'a été observé ni en français, ni en espagnol, où il pourrait même y avoir un mécanisme d'inhibition. À l'inverse, en français et en espagnol, l'existence d'au moins un voisin plus fréquent que le mot cible entraîne un effet inhibiteur.

À nouveau, la problématique se complique dans le cas d'une L2. L'approche monolingue postule que si des effets de voisinages existent, « il est alors possible de considérer que la plupart des mots sont insérés dans un voisinage de mots formellement similaires » (Font et Lavaur, 2004, 378), ou du moins, que l'activation de la représentation du mot cible par le stimulus s'opère en même temps que ses voisins, dont le niveau d'activation influence simultanément la sienne propre. Par conséquent, dans le cas bilingue, la question est de savoir si le stimulus visuel va activer des candidats appartenant à une seule ou à deux langues. Les expériences de Font et Lavaur (2004) révèlent que « la reconnaissance visuelle d'un mot d'une langue est significativement influencée par la proximité orthographique qu'il entretient avec les mots de l'autre langue ». Par conséquent, on doit postuler un accès parallèle et non sélectif aux représentations des deux langues, une hypothèse qui concorde avec le modèle BIA de Van Heuven *et al.* (1998) (*cf.* Chapitre 1).

En conclusion, il apparaît, dans le cas bilingue, que « les liens existants entre les représentations lexicales d'une même langue ne sont pas fondamentalement différents des liens qui existent entre les représentations des deux langues » (Font et Lavaur, 2004, 399) et que les performances à la lecture sont susceptibles d'être influencées à la fois par le voisinage absolu (densité et fréquence des voisins, quelle que soit leur langue) et le voisinage relatif (différences de fréquence et de densité existant entre le voisinage intralangue et interlangue). Il faut aussi noter que cette variable n'a jamais, à notre connaissance, été considérée dans les recherches en lisibilité. Nous avons donc décidé de la retenir parmi nos facteurs d'intérêt.

Mesures de la longueur des mots

Les différents effets lexicaux décrits jusqu'ici présentent l'avantage de constituer fort probablement des causes réelles des difficultés de lecture. Malheureusement, ils sont pour la plupart, assez difficiles à opérationnaliser. Il n'est pas étonnant, dès lors, que les premières et les plus utilisées des variables lexicales aient été des mesures basées sur la longueur des mots, bien plus simples à paramétriser, mais que l'on a longtemps considérées comme des artefacts d'autres variables explicatives. L'hypothèse qui sous-tend cette approche est la suivante :

Hypothèse 3.5 *Plus un mot est long, plus il est susceptible d'être difficile à lire et à comprendre. Par conséquent, plus un texte comprend de mots longs, plus il risque d'être difficile à lire.*

On peut faire remonter cette hypothèse à la thèse de master de Bear, intitulée *The length of Words as an Index of Difficulty in Silent Reading* et publiée en 1927, dont les analyses ont été relayées par Johnson (1930). Bear observe que, pour des textes dont la difficulté est

croissante, le pourcentage de mots polysyllabiques qu'ils contiennent augmente en fonction de la difficulté de ces textes. D'après Klare (1963, 32), elle aurait aussi mis au jour la forte corrélation négative entre la longueur des mots et leur fréquence qui est généralement attribuée à Zipf (1965)[7], laquelle expliquerait en partie pourquoi un mot plus court est plus simple qu'un autre plus long.

Cette hypothèse est ensuite reprise par de nombreux chercheurs en lisibilité (Flesch, 1948; Gray et Leary, 1935; Gunning, 1952; Henry, 1975; Smith, 1961) qui en vérifient empiriquement l'efficacité. Ils recourent à différentes façons de mesurer la longueur des mots, que nous allons envisager dans la suite de cette section. Auparavant, il est intéressant de se demander si réellement les effets de la longueur des mots ne seraient que des artefacts d'autres variables lexicales, comme le pense par exemple Mesnager (1989, 25). Celui-ci, comparant sa propre variable, qui correspond au nombre d'absents de la liste de Catach (1985), au nombre moyen de syllabes employé par Flesch (1948), dit :

> Bien sûr, elle aussi est une approximation[8]. La « vraie » méthode consisterait à affecter tous les mots d'un texte du coefficient de fréquence qui lui est propre [...].

Il semble donc considérer que le nombre de syllabes est moins précis que le nombre d'absents d'une liste de fréquence, lequel ne consiste jamais qu'en une approximation de la fréquence moyenne des mots.

Pourtant, Henry (1969, 227) défendait déjà l'existence d'un effet de la longueur des mots, indépendant de celui de fréquence :

> Les résultats obtenus montrent clairement que les techniques de mesure basées sur la référence à un vocabulaire de base quelconque conduisent à des résultats pratiquement identiques. [...] Par contre, il est tout aussi évident que la mesure de la longueur moyenne des mots est représentative d'un phénomène différent. Certaines indications font penser que cette mesure pourrait être représentative du niveau d'abstraction.

Cette intuition a depuis été confirmée par quelques études en psychologie cognitive de la lecture. Tout d'abord, il faut rappeler une donnée déjà évoquée dans le Chapitre 1, à savoir que la qualité de vision se dégrade à mesure qu'un symbole à reconnaître s'éloigne du point fixé par la fovéa. Or, des études ont montré qu'il existerait un endroit optimal où fixer le regard lorsqu'on cherche à reconnaître un mot et que le temps de reconnaissance de ce mot

7. Signalons que Mesnager (1989, 22) a lui aussi obtenu une corrélation entre la fréquence des mots et leur nombre de lettres pour les 2 000 mots les plus fréquents de la langue française, tirés de la liste de Catach (1985).

8. Il parle ici de sa variable.

augmenterait d'environ 20ms par lettre d'écart par rapport à ce point optimal (O'Regan et Jacobs, 1992; Vitu *et al.*, 1990, 591). En effet, lorsque le regard ne s'est pas posé sur ce point idéal, il s'avère parfois nécessaire de le repositionner, ce qui allonge le temps nécessaire à la reconnaissance du mot. Vitu *et al.* (1990) montrent que, pour les mots plus longs, il existe davantage de points de fixation non optimaux – situés en particulier au début et à la fin du mot – et que la probabilité d'une telle refixation augmente. C'est probablement ce qui explique que O'Regan et Jacobs (1992) obtiennent un effet de la longueur des mots sur le temps nécessaire à la réalisation de tâches de prononciation et de décision lexicale ; effet qui est bien distinct de l'effet de fréquence.

Autrement dit, la longueur des mots ne serait pas seulement un artefact d'autres causes de la difficulté lexicales, mais influencerait également la reconnaissance des mots. Le succès de cette variable nous semble dès lors amplement mérité. Elle fût en effet l'un des facteurs les plus utilisés en lisibilité, sous diverses formes que différencie l'entité linguistique utilisée comme unité de mesure : tantôt la syllabe, tantôt la lettre, tantôt encore l'affixe. Nous allons brièvement revenir sur chacune d'elles, afin d'en déterminer les atouts et les défauts.

Les mesures basées sur les syllabes Étonnamment, il s'agit de la première unité utilisée en lisibilité pour évaluer la difficulté des mots. Pourtant, elle constitue a priori un choix moins évident que la lettre, notamment parce qu'elle est plus difficile à délimiter précisément que cette dernière. Ce choix s'expliquerait par le fait que l'anglais était (et reste) la principale langue d'intérêt en lisibilité. Or, par rapport au français, l'anglais, par exemple, comprend un plus grand nombre de mots monosyllabiques, qui appartiennent principalement au fond germanique et sont plus utilisés dans la vie quotidienne (Damerst, 1972, 63). Dès lors, il est sans doute apparu plus pratique d'employer le pourcentage de mots comportant plus d'une ou deux syllabes que de calculer le nombre moyen de lettres par mot.

Ainsi, on trouve des auteurs ayant expérimenté le pourcentage de mots polysyllabiques (Dolch, 1948; Johnson, 1930; Wheeler et Smith, 1954) ou, à l'inverse, le pourcentage de mots monosyllabiques (Caylor *et al.*, 1973; Dale et Tyler, 1934; Damerst, 1972; Farr *et al.*, 1951), voire les deux (Gray et Leary, 1935). À chaque fois, cette variable s'est révélée significativement corrélée avec la difficulté des textes. D'autres chercheurs ont considéré que les mots difficiles étaient mieux circonscrits à l'aide du pourcentage de mots de plus de deux syllabes (Caylor *et al.*, 1973; Feng *et al.*, 2009; Gunning, 1952; McLaughlin, 1969), reculant de ce fait le seuil de la difficulté. Enfin, la mesure qui a connu le plus de succès est le nombre moyen de syllabes par mot, popularisé par Flesch (1948) et repris par de nombreux autres expérimentateurs (Bormuth, 1966; Caylor *et al.*, 1973; Feng *et al.*, 2009; Fry, 1963; Kincaid *et al.*, 1975; Schwarm et Ostendorf, 2005).

Derrière ces différentes facettes du même phénomène, se cache en vérité la question du

pouvoir discriminant des variables ainsi définies, que nous avons déjà abordée au sujet du pourcentage de mots absents d'une liste de référence. De la même façon que le pouvoir discriminant de cette variable est déterminé par la taille de la liste, le choix de considérer tous les mots polysyllabiques comme difficiles, ou seulement ceux comportant plus de trois syllabes est lourd de conséquences. À notre connaissance, aucune étude n'a été capable de faire l'unanimité autour d'un seuil optimal. Par ailleurs, un autre défaut des mesures basées sur les syllabes a été mis en évidence par Smith (1961). Il considère en effet que le comptage des syllabes manque de fiabilité par rapport au comptage des lettres et plaide en faveur de cette autre unité que nous allons maintenant envisager.

Les mesures basées sur les lettres En ce qui concerne les lettres, McClusky (1934) est probablement le premier a avoir employé leur moyenne par mots comme indice de lisibilité. Bien qu'il ait attesté une variation sensible en fonction des sujets des textes de son échantillon, son étude ne fait pas florès. Il faut attendre Bloomer (1959) et surtout Smith (1961) pour que le nombre moyen de lettres soit à nouveau envisagé comme un prédicteur en lisibilité. Smith effectue deux expériences qui tendent à confirmer que cette variable basée sur la lettre peut validement remplacer les mesures de longueur reposant sur la syllabe ou l'affixe. Il analyse d'abord la corrélation entre le nombre de lettres et la fréquence (liste de Thorndike) et observe que, jusqu'à un certain niveau de fréquence, un mot plus rare tend à être plus long. Ensuite, il compare le nombre de syllabes et le nombre de lettres et remarque que ces deux aspects sont bien corrélés.

Par la suite, d'autres études testent ce facteur, aussi bien pour la L1 (Bormuth, 1966, 1969a; Carver, 1974a; Coleman et Liau, 1975; Richaudeau, 1973; Spaulding, 1951) que pour la L2 (François, 2009a; Greenfield, 2004). Sa principale qualité est la simplicité de ses normes de comptage, qui le rend plus fiable que les mesures basées sur la syllabe et surtout l'affixe, aussi bien manuellement (Harrisson, 1980, 57) qu'automatiquement (Smith, 1961). De plus, sa capacité prédictive est également élevée, comme le montre le r de $-0,721$ obtenu par Bormuth (1969a).

Parallèlement à cette paramétrisation basée sur la moyenne, quelques chercheurs ont adopté une autre position, sensiblement similaire à celle qui consiste à n'envisager que les mots de plus de x syllabes. Il s'agit de ne considérer comme difficiles que les mots de plus de x lettres. Sa première attestation est attribuée à Barker et Stokes (1968), qui testent pour x les valeurs comprises entre 6 et 10 et obtiennent une corrélation de $0,78$ entre ce facteur (pour $x = 6$) et le nombre d'absents de la liste de Dale (3000 mots). À la même époque, Björnsson évalue ce prédicteur dans des articles en suédois (Björnsson, 1983) et il est surprenant de voir qu'il parvient de façon indépendante à la même valeur pour x que Barker et Stokes (1968). Celle-ci est encore utilisée par Caylor *et al.* (1973).

Entre ces deux approches, la moyenne ou un seuil, il n'est pas évident de déterminer ce qui convient le mieux à la lisibilité du FLE. Signalons que Harris et Jacobson (1974, 2), qui considèrent comme difficiles les mots comportant plus de 5 lettres, rapportent que cette variable possède un plus grand pouvoir prédictif que le nombre moyen de lettres par mots. Toutefois, ils ont travaillé sur l'anglais et en L1. Or, les mots français sont en moyenne légèrement plus longs que les mots anglais (Kandel et Moles, 1958). C'est pourquoi, il nous est apparu nécessaire d'effectuer de nouvelles expérimentations en vue de comparer ces deux mesures et de fixer un seuil adéquat pour le FLE.

Les mesures basées sur les affixes Beaucoup plus rarement utilisé, le nombre d'affixes par mot a été proposé par Flesch (1943, cité par Chall, 1958), mais a fait l'objet de nombreuses critiques. Flesch (1948) lui-même ne tarde pas à critiquer cette variable qui se révèle pénible à utiliser et peu fiable, car deux personnes analysant le même passage risquent fort de se retrouver avec un décompte final divergeant (Dale et Chall, 1948a, 12). Elle a été réutilisée par Aquino *et al.* (1969), qui ont étudié la proportion de suffixes d'origine latine et ont découvert qu'il s'agit d'un bon prédicteur des performances à un test de closure. Malgré cela, nous ne l'avons pas retenue pour nos expérimentations.

Conclusion : Il nous apparaît que les mesures de la longueur du mot sont pertinentes dans le contexte de la lisibilité en L2. D'une part, parce qu'elles ont déjà fait leurs preuves de manière empirique : Uitdenbogerd (2003) utilise le nombre de mots de plus de deux syllabes, tandis que le nombre moyen de lettres par mot s'est révélé un bon prédicteur chez Greenfield (2004) et François (2009a). D'autre part, puisque l'effet de longueur semble dû à des particularités de la perception visuelle, on peut au moins s'attendre à ce qu'il agisse de manière similaire chez des apprenants dont l'alphabet est aussi latin. D'ailleurs, Cole (1982, cité dans Laufer, 1997, 144) repère chez des apprenants un fort effet de la longueur des mots sur la reconnaissance de ceux-ci. Apparemment, cet effet est particulièrement évident avec l'apprenant dont la L1 emploie un alphabet non latin, ce qui renforce notre conviction que les mesures de la longueur du mot sont adaptées à la lisibilité du FLE.

Il faut toutefois relayer la critique pertinente de Laroche (1979, 133) à l'encontre de ce type de mesure et qui concerne une fois de plus la question des congénères. Il souligne qu'une proportion importante des termes anglais transparents pour un natif francophone sont plutôt longs, car il s'agit d'emprunts « latins ». Dès lors, l'association entre longueur et difficulté en devient faussée. Laroche (1979) estime pour sa part que la longueur des mots n'est un facteur valide que pour une paire de langues où le taux de congénères est très réduit.

Conclusion

Suite à cet inventaire critique des variables lexicales utilisables en lisibilité du FLE, nous remarquons que chacune des familles envisagées, à l'exception des mesures basées sur l'âge d'acquisition, est non seulement apparue appropriée pour prédire la difficulté de textes, mais qu'un effet spécifique a pû être détecté expérimentalement, une fois certains facteurs proches contrôlés. A priori, cela signifie que ces facteurs constituent non seulement de bons indices de lisibilité, mais sont aussi des causes réelles des difficultés de compréhension à la lecture.

Toutefois, il faut rester prudent face à ce tableau optimiste. D'une part, nous n'avons pas connaissance d'études contrôlant l'ensemble des facteurs que nous avons décrits. Rien ne nous dit que certains des effets détectés ne sont pas dus en réalité à une autre variable non contrôlée. D'autre part, il n'est pas évident d'évaluer la **taille** (c'est-à-dire l'importance) de l'effet de chacune de ces variables lorsqu'elles agissent ensemble, puisque les expérimentations visent en général à isoler l'un ou l'autre facteur. Il existe évidemment divers modèles, tels que le modèle à deux voies ou les approches connexionnistes (*cf.* Chapitre 1), qui combinent ces facteurs en un modèle prédictif. Cependant, il s'agit de simulations dont la validité est jugée à l'aune de leur capacité à reproduire les résultats d'expériences précédentes.

Par conséquent, il nous semble actuellement impossible de hiérarchiser les facteurs lexicaux par ordre d'importance sur la base de ces études théoriques. Nous en resterons donc à l'approche classique qui consiste à juger leur efficacité en fonction de leur niveau d'association avec le critère utilisé, une façon de faire que Chall (1958, 45) critique ainsi : « la supériorité de l'une ou l'autre [variable] dépendrait de la nature du critère et de la fiabilité de son emploi ». Qu'il soit donc clair que notre étude n'aura pas comme objectif de poser un jugement quant à la supériorité universelle d'une variable sur une autre, comme certains chercheurs en lisibilité ont cru pouvoir le faire, mais uniquement d'observer leur efficacité dans le cadre de notre contexte d'expérimentation.

3.2.3 Les variables au niveau syntaxique

Le niveau syntaxique est historiquement le deuxième à avoir été appréhendé en lisibilité : dès Vogel et Washburne (1928), on recourt à des variables de surface visant à estimer la complexité des structures présentes dans les textes. Contrairement aux facteurs lexicaux, elles ne concernent pas le décodage des mots, mais ressortissent directement de l'étape de compréhension (Rayner *et al.*, 2001, 49), que nous avons déjà traitée dans le Chapitre 1. En parcourant la littérature, nous avons identifié quatre classes de mesures syntaxiques, qui s'attachent respectivement à la longueur des phrases, la complexité des structures, les formes verbales et les catégories grammaticales. Cette section les passe en revue, cherchant toujours

à en justifier la pertinence à l'aide de données empiriques portant sur les mécanismes de lecture en L1 et – quand elles existent –, en L2.

Soulignons que les deux premières familles représentent de loin les mesures les plus employées dans la discipline et que les avis divergent quant à leur relation. Pour Bormuth (1966, 122), la longueur des phrases serait relativement indépendante de la complexité syntaxique, du moins telle qu'il la mesure. Il y aurait donc motif à combiner ces deux variables au sein d'une même formule. Toutefois, Smith (1961) et Kemper *et al.* (1993, 418) sont convaincus du contraire et s'accordent à dire qu'une augmentation de la longueur des phrases s'effectue généralement via l'inclusion de plusieurs propositions subordonnées, lesquelles augmentent la complexité syntaxique de la phrase. Cette question essentielle est discutée en détail dans la section suivante.

Mesures de la longueur des phrases

À notre connaissance, les premiers à avoir envisagé une mesure de la longueur phrastique en lisibilité sont Vogel et Washburne (1928), qui lui préfèrent toutefois une mesure de la complexité syntaxique pour leur formule finale. Ils notent une nette association entre ces deux types de mesures : dans un échantillon de 75 phrases, le nombre de mots est fortement corrélé au nombre de phrases simples ($r = -0,741$), et plus encore au nombre de syntagmes adverbiaux ($r = 0,818$). Cette tendance à tester les deux types de facteurs, mais à ne retenir que celui qui recouvre la complexité syntaxique se retrouve chez Dale et Tyler (1934).

Cette préférence pour les mesures de la complexité syntaxique dans les premières recherches en lisibilité s'explique probablement par l'opinion selon laquelle la longueur des phrases ne serait qu'un indice de la complexité. Ainsi, Dale et Tyler (1934, 394) considèrent plausible que « soient plus difficiles des phrases qui impliquent une suspension du jugement jusqu'à ce que la phrase entière ait été appréhendée. Il pourrait alors apparaître que la longueur des phrases soit corrélée positivement avec la difficulté [...] ».

C'est durant la période classique, où la mode est à la simplicité, que les mesures de la longueur phrastique prennent l'ascendant sur celles de la complexité structurelle. En effet, elles sont plus simples à calculer pour un néophyte, lequel n'a pas besoin de pouvoir reconnaître une phrase complexe ou un syntagme adverbial. Privilégié par Flesch (1948) et Dale et Chall (1948a), le nombre moyen de mots par phrase connaît ensuite un large succès auprès des chercheurs en lisibilité. Il est vrai qu'au-delà de son aspect pratique, il reste insolemment efficace, au point que Chall et Dale (1995, 5) affirment : « la longueur des phrases reste tout à fait valide comme prédicteur de la complexité syntaxique, et même davantage que des mesures plus compliquées reposant sur des théories linguistiques sophistiquées (*cf.*

les études de Bormuth, 1971, et de MacGinitie et Tretiak, 1971) ».

Justifications : Au-delà de l'aspect pratique et de l'efficacité empirique des variables basées sur la longueur des phrases, nous avons voulu déterminer si leurs effets sur le processus de la lecture en L1 et en L2 sont attestés par des expérimentations. Cependant, les études actuelles sur le traitement de la syntaxe portent plutôt sur les structures syntaxiques que sur l'influence de la longueur de la phrase. Cette thématique semble d'ailleurs délaissée depuis un certain temps, ce qui nous laisse penser qu'on l'a estimée peu pertinente comme cause de la difficulté. L'étude de Schlesinger (1968, 71-80) va d'ailleurs dans ce sens. L'auteur mesure le temps de lecture et le taux de compréhension de 58 sujets sur deux séries de 3 fragments dont la longueur des phrases varie, mais dont la nature des structures syntaxiques et le taux de redondance est contrôlé. Il n'obtient pas d'effets spécifiques significatifs de la longueur des phrases en hébreu.

Pourtant, plus récemment, Labasse (1999) défend l'influence de la longueur phrastique sur la lecture à l'aide de trois arguments. Le premier d'entre eux s'inspire des travaux de Chomsky et se résume ainsi : « la compréhension d'une phrase impliquerait la reconstitution de sa structure syntaxique : plus la phrase sera longue, donc compliquée, plus il sera difficile d'en reconstruire l'arborescence grammaticale » (Labasse, 1999, 89). Autrement dit, l'auteur en revient à la conception que la longueur phrastique ne constitue qu'un artefact de la complexité syntaxique, qu'elle peut estimer efficacement.

Ses deux autres allégations défendent par contre l'idée d'un effet spécifique de la longueur. Ainsi, Labasse (1999, 90) met en relation les limites de l'empan mnésique décelées par Miller (1956) avec le traitement de la phrase. Si le nombre de mots excède les capacités limitées de la mémoire de travail, le lecteur ne serait pas capable de mémoriser une phrase dans son ensemble et, pour autant que l'on accepte l'association entre taux de mémorisation et compréhension, il faut en conclure qu'une phrase plus longue serait plus difficile à appréhender. Cette hypothèse est également défendue par Graesser *et al.* (2004, 194). Pourtant, encore faut-il que le cerveau se serve réellement de la phrase comme unité de traitement, ce qui est loin d'être avéré.

En effet, Richaudeau (1969) estime, sur la base d'expériences de mémorisation, que le lecteur retient plutôt des sous-phrases. Dès lors, la phrase pourrait bien ne pas être l'unité la plus adaptée pour évaluer les difficultés de mémorisation à la lecture. Les expériences de Coleman (1962, cité par Bormuth, 1966, 93) vont dans le même sens ; il estime que la proposition indépendante constitue une unité plus valable que la phrase canonique. Peu après, Bormuth (1969a, 53) lui-même obtient un meilleur coefficient de corrélation pour les « unités minimales de ponctuation » que pour les phrases. Il en conclut que l'effet de la longueur d'une phrase s'explique davantage par les structures que celle-ci contient que simplement

par son nombre de mots.

D'autres résultats corroborent cette hypothèse. Ainsi, Kolers (1970, 108) observe, au terme de plusieurs expériences basées sur des textes géométriquement transformés, que le nombre d'erreurs augmente avec la complexité des phrases, mais pas simplement en fonction du nombre de mots dans une proposition. Rappelons enfin les travaux de Granowsky et Botel (1974) et de Davison et Kantor (1982) (*cf.* Chapitre 2) qui s'attaquent précisément à l'utilisation de la longueur des phrases en lisibilité et mettent en évidence que manipuler celle-ci ne conduit pas nécessairement à un texte plus simple. Blau (1982) parvient à la même conclusion pour des apprenants de l'anglais L2.

La troisième justification proposée par Labasse (1999) fait référence au modèle de Kintsch et van Dijk (1978), considérant que l'analyse et la structuration propositionnelle qui seraient effectuées dans l'esprit du lecteur s'accordent mieux à des phrases courtes. Toutefois, Kintsch et van Dijk (1978, 379) reconnaissent eux-mêmes qu'ils ont émis l'hypothèse que chaque cycle de restructuration de la structure propositionnelle stockée en mémoire s'effectue phrase par phrase sans preuve en ce sens et pour des raisons d'illustrations. Or, les études précitées laissent penser qu'il est plus vraisemblable que d'autres unités syntaxiques soient utilisées.

En conclusion, il nous apparaît que la longueur de la phrase est très probablement un intermédiaire de la complexité syntaxique et qu'elle n'apporte pas une information propre. Cela n'empêche pas cette variable de rester l'un des indices syntaxiques les plus efficaces pour la lisibilité de textes en L1 comme en L2, même si elle présente un défaut grave : si l'on prend les mots d'un texte et que l'on les mélange au sein de chaque phrase, on obtiendra le même score de lisibilité pour ce texte.

Les différentes mesures de la longueur phrastique La mesure la plus commune de la longueur phrastique est bien entendu le nombre moyen de mots par phrase, employés par de nombreux chercheurs (Bormuth, 1966; Caylor *et al.*, 1973; Chall et Dale, 1995; Farr *et al.*, 1951; Feng *et al.*, 2009; Gunning, 1952; Harris et Jacobson, 1974; Henry, 1975; Kincaid *et al.*, 1975; Schwarm et Ostendorf, 2005; Smith, 1961; Stenner, 1996; Uitdenbogerd, 2003). Il existe cependant d'autres approches, car la moyenne présente un défaut dont nous avons déjà parlé à propos de la fréquence. Elle ne considère pas la variabilité, alors que la présence de phénomènes extrêmes est susceptible d'engendrer des problèmes de compréhension, en particulier dans le cas du FLE.

C'est pourquoi apparaît chez certains chercheurs la volonté de se focaliser sur les observations les plus problématiques pour la compréhension. Dolch (1948, 246) propose de se baser sur la distribution du nombre de mots par phrase pour un texte et d'utiliser la différence entre

la médiane et ce qu'il appelle « la phrase longue typique », laquelle correspond au 9e décile de cette distribution. De leur côté, MacGinitie et Tretiak (1971) se limitent à employer le 75e percentile, tandis que Daoust *et al.* (1996, 217) testent le pourcentage de phrases de plus de 30 mots, qui se révèle être leur meilleure variable, plus efficace notamment que la proportion de points, aussi employés par ces auteurs québécois.

Parallèlement à ces efforts, germe l'idée qu'un texte constitué de phrases trop courtes représente aussi une difficulté. Déjà envisagée par Schlesinger (1968, 79), elle est reprise par Damerst (1972), qui estime que la difficulté et le nombre de mots par phrase entretiennent une relation parabolique. Il existerait un intervalle de valeurs optimales (de 14 à 17 mots), tandis que les phrases plus longues et plus courtes nuiraient à une lecture efficace. Richaudeau (1984, 184) abonde dans le même sens : « Un texte trop segmenté en très courtes phrases est moins bien retenu qu'un texte équivalent écrit en phrases moyennes. [..] Les coupures trop fréquentes entre les suites de mots « cassent » le fil conducteur de l'argumentation, du sens, et rendent ainsi impossibles les anticipations syntaxiques ou sémantiques de la part du lecteur ».

Malgré l'intérêt de cette dernière hypothèse, elle ne nous paraît pas indiquée au cas du FLE, où les premiers textes proposés à l'apprenant comportent des phrases très courtes, puisqu'il ne maîtrise pas encore le système syntaxique de sa L2. Le manque de cohésion apparaît secondaire dans ce cas. Par conséquent, dans cette étude, nous nous sommes cantonné à la mesure la plus classique : le nombre moyen de mots par phrase.

Mesures de la complexité syntaxique

Comme nous l'avons dit précédemment, les mesures essayant de refléter la complexité des structures présentes dans un texte ont également été envisagées pour la première fois par Vogel et Washburne (1928). Ceux-ci ont ainsi défini plusieurs variables superficielles assez simples telles que le nombre de phrases simples parmi 75, le nombre de syntagmes adverbiaux dans 75 phrases, etc. Largement utilisées dans les premiers travaux en lisibilité, les mesures de la complexité syntaxique vont connaître un déclin progressif à cause de la complexité de leur paramétrisation, avant que les études syntaxiques de Chomsky et d'autres les remettent à l'ordre du jour.

Dans la lignée des grammaires génératives et transformationnelles, plusieurs études sur la difficulté des structures syntaxiques sont publiées dans les années 60 et 70 (Bormuth, 1969a; Henry, 1975; MacGinitie et Tretiak, 1971; Schlesinger, 1968). Elles démontrent que certaines structures grammaticales sont plus difficiles à comprendre et cherchent à expliquer ce phénomène à l'aide de concepts issus de la Grammaire Générative et Transformationnelle

(GGT) Chomsky (1957, 1965), tels que le nombre de transformations nécessaires pour passer de la structure profonde de la phrase à la structure de surface ou le niveau d'enchâssement. Il s'agit là de variables plus complexes que les variables syntaxiques superficielles précédentes, qui se révèlent malheureusement rarement plus efficaces. De plus, si ces expérimentations détectent des effets de la complexité syntaxique, elles peinent à les expliquer en relation avec la GGT.

Par la suite, l'impact des structures syntaxiques sur la lisibilité des textes est délaissé au profit des facteurs cognitifs, sans doute aussi parce qu'il est difficile de concevoir des tâches expérimentales qui ne mobilisent que des processus syntaxiques. C'est avec l'essor des techniques d'analyse des processus cérébraux (*cf.* Chapitre 1) qu'apparaissent de nouvelles hypothèses explicatives de la difficulté syntaxique. Kutas et Hillyard (1980) sont parmi les premiers à aborder la question de la complexité syntaxique avec une nouvelle technique, les PEV qui mesurent l'activité électrique du cerveau à l'aide d'électrodes. Cette technique permet d'observer directement le processus de lecture sans passer par ses produits. Les signaux PEV sont constitués de plusieurs composants, positifs (P) ou négatifs (N), situés temporellement en référence au moment où est présenté un stimulus. Ainsi, le composant détecté par Kutas et Hillyard (1980), qu'ils appellent N400, est une onde négative qui apparaît autour des 300 ms, dont l'amplitude culmine à 400 ms et perdure jusqu'à 600 ms.

Trois composants principaux ont été détectés chez des natifs lors de la lecture de phrases problématiques. Le premier, l'Early Left Anterior Negativity (ELAN), est un composant négatif qui apparaît dans la partie antérieure gauche du cerveau entre 150 ms et 250 ms à la lecture de phrases contenant une violation syntaxique. Hahne et Friederici (1999) l'associent à une première passe du traitement syntaxique, hautement automatisée. Il est suivi par le N400, typiquement provoqué par un mot sémantiquement incongru, que Kutas et Hillyard (1980) interprètent comme un signe d'une difficulté d'intégration sémantique. Ce serait en effet à ce moment que sont attribués les rôles thématiques (Hahne et Friederici, 1999, 195). Plus précisément, Kutas et al. (1984, cités par Kounios et Holcomb (1994, 806)) ont montré que l'amplitude de N400 est une fonction monotonique de la probabilité de closure des mots dans les phrases, autrement dit, l'amplitude de N400 est plus large pour les mots peu prédictibles d'après le contexte.

Enfin, lorsque les informations syntaxiques et sémantiques n'ont pas été correctement intégrées au cours des deux phases précédentes, on observe l'apparition d'un troisième composant, le P600, qui signale l'existence d'une phase de réanalyse syntaxique, moins automatisée. Celui-ci a d'abord été associé à la lecture de mots agrammaticaux ou incongrus en fonction du contexte syntaxique précédent [9] (Osterhout et Holcomb, 1992). Toutefois,

9. C'est le cas au sein de phrases appelées « garden-path sentences » : il s'agit d'un cas où le lecteur, après

Kaan *et al.* (2000) ont montré que des phrases complexes peuvent aussi engendrer un P600. Dès lors, celui-ci pourrait plutôt être interprété comme le signe d'une difficulté d'intégration syntaxique que comme la trace d'un processus de réparation ou de réanalyse.

Ces découvertes aident à mieux comprendre la question de la complexité syntaxique. D'une part, elles attestent de l'existence de processus cérébraux différents entre phrases plus simples et plus complexes. Kaan *et al.* (2000, 161) soulignent ainsi que l'amplitude de P600 augmente d'ailleurs en fonction de la difficulté qu'il y a à construire la représentation syntaxique d'une phrase donnée. D'autre part, il apparaît que le traitement des phrases syntaxiquement complexes allonge le temps de traitement des mots problématiques, employant des ressources qui auraient pu être utilisées pour poursuivre la lecture. On en arrive ainsi à l'explication principale de l'effet de la complexité syntaxique sur la compréhension à la lecture : les limitations mémorielles.

En effet, Crain et Shankweiler (1988) ont effectué une expérience montrant que les problèmes de lecture liés à la syntaxe ne s'expliquent pas parce que des jeunes lecteurs ne connaissent pas les structures incriminées, mais plutôt parce que celles-ci sont dues aux limitations de la mémoire de travail à gérer la complexité. Cette observation empirique nous semble être éclaircie par le fait que des phrases syntaxiquement difficiles provoquent l'apparition d'un composant P600. Or, celui-ci constituerait un processus moins automatisé que l'ELAN, et donc plus gourmand en ressources mentales (Hahne et Friederici, 1999, 195).

La situation en L2 : Dans le cas d'un apprenant de L2, la situation est plus problématique, puisque les structures syntaxiques ne préexistent pas à l'apprentissage de la lecture, comme chez les natifs. Il lui faut donc pratiquer longuement cette L2 avant d'atteindre une certaine automaticité au niveau des processus d'analyse syntaxique. Nous avons vu que, chez les natifs, le composant ELAN trahit la présence de processus automatisés. Or, il semblerait que ce composant ne soit pas présent chez des apprenants de L2, sauf sous certaines conditions spécifiques.

C'est Weber-Fox et Neville (1996) qui ont employé en premier les PEV pour étudier le traitement syntaxique de l'anglais chez des apprenants chinois adultes, mais ayant acquis l'anglais à différents âges. Leurs résultats montrent un profil très différent entre natifs et apprenants d'une L2. Aucun de ces derniers ne présente d'ELAN, même les bilingues précoces (de 1 à 3 ans). Ce résultat a notamment été répliqué par Hahne (2001) sur des apprenants russophones de l'allemand. Toutefois, Isel (2007) parvient à susciter ce composant chez des sujets germanophones ayant acquis tardivement le français, lorsque ceux-ci sont confron-

analyse d'une première partie de la phrase, s'attend à une rencontrer une structure syntaxique spécifique, mais doit revoir son jugement lorsqu'il est confronté à la structure réelle qui diffère de celle qu'il avait imaginé (Crain et Shankweiler, 1988, 215)

tés à une rupture syntaxique concernant un constituant obligatoire de la phrase, et non une violation plus mineure.

Suite à ces expériences, van Hell et Tokowicz (2010, 59) concluent qu'une première phase automatique d'analyse syntaxique apparaît parfois chez des lecteurs en L2, lorsqu'ils sont confrontés à des structures simples ou une rupture syntaxique grave. Ils interprètent ces résultats selon le modèle compétitif de MacWhinney (1997), estimant que l'absence d'un ELAN s'explique par la compétition entre les deux langues pour l'interprétation des indices. Plus la L1 est éloignée de la L2 et plus la compétition entre les deux systèmes sera forte. Par contre, plus les indices sont importants ou plus l'âge d'acquisition de la L2 est précoce, et moins joue cette compétition. Malgré tout, on note que les processus d'analyse syntaxique des apprenants sont moins rapides et automatisés que chez des natifs, même chez les bilingues précoces.

Par ailleurs, il semblerait aussi que la tâche de ré-analyse, matérialisée par la présence d'un composant P600, soit également problématique chez les lecteurs en L2. Ainsi, Weber-Fox et Neville (1996) observent que si les bilingues relativement précoces (ayant commencé l'acquisition de la L2 avant 10 ans) présentent un tel composant, celui-ci est retardé chez les 11 à 13 ans et n'apparaît pas chez les adultes ayant commencé l'apprentissage après 16 ans (du moins chez leurs sujets bilingues anglais-chinois). Parallèlement, ces apprenants sont moins précis que des monolingues dans leurs jugements sur la grammaticalité des phrases qui leur sont proposées, en particulier ceux ayant acquis leur L2 tardivement. On notera que cet effet de l'âge d'exposition joue nettement moins en ce qui concerne les jugements sémantiques, pour lesquels seuls les sujets tardifs (après 16 ans) éprouvent des difficultés. Ces résultats révèlent que les apprenants de L2 sont moins sensibles aux ruptures syntaxiques que sémantiques, et sont également moins capables de réparer celles-ci. Weber-Fox et Neville (1996, 248) estiment ainsi qu'« une exposition précoce pourrait être nécessaire pour atteindre une compétence grammaticale équivalente à celle de natifs ». Ces observations ont été confirmées entre autres par Hahne (2001), qui obtient un P600 retardé de 150 ms pour des apprenants avancés.

En résumé, ces travaux tracent le portrait d'un lecteur en L2 moins rapide et moins précis dans son analyse des constructions syntaxiques. Il subit des interférences avec sa L1, qui sont bénéfiques lorsque les structures de la L2 existent aussi dans sa L1, mais le freine dans le cas contraire (Bernhardt, 1991, citée par Bernhardt et Kamil, 2006, 91. Il doit ainsi souvent mobiliser plus de ressources mentales pour l'analyse syntaxique qu'un natif. Bernhardt et Kamil (2006, 91) défendent d'ailleurs le fait que la syntaxe soit une variable clef pour prédire la compréhension à la lecture en L2. C'est pourquoi, nous envisageons ici différentes mesures visant à prendre en compte la complexité syntaxique dans une formule de lisibilité.

Nous les avons regroupées en six catégories, selon qu'elles soient basées sur la proportion de phrases simples, les types de propositions, les structures syntaxiques, divers concepts issus du modèle transformationnel, la notion de poids ou la redondance des structures.

A. La proportion de phrases simples : Ce premier facteur repose sur un postulat simple : plus un texte comprend de phrases simples, plus il est abordable. Une fois de plus, ce sont Vogel et Washburne (1928) qui, les premiers, ont envisagé de calculer la proportion de phrases simples, c'est-à-dire de phrases constituées d'une seule proposition indépendante, parmi un échantillon de 75. Bien que cette variable n'apparaît pas très efficace (sa corrélation avec le critère est de $-0,371$), les auteurs l'intègrent dans leur formule. Peu après, Dale et Tyler (1934) distinguent les phrases simples, composées (qui contiennent deux propositions indépendantes coordonnées) et complexes (qui comprennent une proposition indépendante associée à une ou plusieurs propositions subordonnées). Ils obtiennent des résultats moins encourageants encore : $r = 0,146$ pour le ratio de phrases simples et $r = -0,149$ pour la proportion de phrases complexes et composées.

À l'opposé de ces premières expériences peu concluantes, Ojemann (1934) obtient une corrélation de $-0,61$ pour la proportion de phrases simples, tandis que Washburne et Morphett (1938) incluent ce facteur dans leur formule. Pourtant, celui-ci n'est plus guère utilisé par la suite, cédant le pas à d'autres facteurs. À notre connaissance, il n'a pas non plus été utilisé en lisibilité des L2, si ce n'est par Stone et Mullins (1932). Pourtant, les résultats des analyses PEV révèlent que des phrases actives simples provoquent un ELAN, mais pas les phrases plus complexes. En conséquence, ce ratio nous semblerait adapté au cas du FLE.

B. Types de syntagmes et de propositions : Se limiter à la proportion de phrases simples reste une approche rudimentaire de la complexité syntaxique, puisqu'une phrase complexe peut correspondre à des structures de complexité très variée. Dès lors, Vogel et Washburne (1928) explorent une autre voie : ils prennent en considération le nombre de syntagmes ($r = 0,576$), mais surtout, ils se focalisent sur certains types de syntagmes et propositions, tels que les adverbiaux ($r = 0,518$) et les adjectivaux ($r = 0,518$). Cette variable, dont les règles de dénombrement ne nous semblent pas si simples, connaît pourtant un large succès. Les chercheurs s'intéressent au nombre de syntagmes prépositionnels (Bormuth, 1966; Dale et Tyler, 1934; Gray et Leary, 1935; Lorge, 1944a), au nombre de relatives et de subordonnées non temporelles (Henry, 1975) ou encore au nombre de propositions indéterminées (Dale et Tyler, 1934). Bormuth (1966) propose par ailleurs une variante unique : la proportion de mots compris dans des syntagmes propositionnels.

Plus récemment, cette famille de variables a fait l'objet de procédures d'automatisation. Schwarm et Ostendorf (2005) et Petersen et Ostendorf (2009) recourent à un analyseur syntaxique automatique pour détecter le nombre moyen de syntagmes nominaux et verbaux,

ainsi que de « SBAR »[10]. Les auteurs concluent à la moindre efficacité de ce type de prédicteurs par rapport à des facteurs lexicaux, mais insistent sur le fait qu'ils apportent une information supplémentaire utile. Par contre, Pitler et Nenkova (2008) rapportent des corrélations non significatives pour ces mêmes indices, à l'exception des syntagmes verbaux (VP, $r = 0,42$). Quant à Feng *et al.* (2010), ils comparent les performances de classifieurs SVM basés uniquement sur chacune de ces variables, ainsi que le nombre moyen de syntagmes propositionnels. Il en ressort que le nombre moyen de syntagmes nominaux et le nombre moyen de syntagmes verbaux seraient les meilleurs prédicteurs parmi ce type de variables.

C. Les structures syntaxiques : Un moyen d'aller plus loin que le simple repérage de syntagmes ou de propositions particulières est de dénombrer certaines structures particulières, supposées responsables des difficultés d'analyse syntaxique chez le lecteur. Cette stratégie est initiée par Bormuth (1969a), qui liste 50 structures syntaxiques particulières et définit autant de variables correspondant simplement au nombre de structures de chaque type. Il obtient pour ces variables un certain nombre de corrélations significatives, dont la meilleure atteint un r de $-0,456$. Par la suite, Bormuth *et al.* (1970) testent la moitié de ces constructions sur des élèves de 4e primaire en vue de les hiérarchiser d'après leur complexité, évaluée selon le taux de bonnes réponses à des questions de compréhension portant sur les phrases à évaluer. Notons que pour le français, Henry (1975) identifie une dizaine de constructions qu'il considère complexes, telles que les infinitives, des complétives, des réciproques et réfléchies, etc.

Cette approche présente toutefois plusieurs défauts. Tout d'abord, elle accroît considérablement le nombre de variables. Certaines structures, pourtant compliquées, sont si rares que leur corrélation avec le critère apparaît non significative. De plus, les efforts requis pour les paramétriser manuellement sont exagérés par rapport à ce qu'elles apportent comme information. Aujourd'hui, on peut cependant recourir à un analyseur syntaxique automatique, capable de repérer ces structures. Cette façon de faire a été appliquée à l'ALS par Heilman *et al.* (2007). Ceux-ci détectent 22 constructions syntaxiques identifiées à partir de livres de grammaire (ensemble 1) et évaluent leur pouvoir prédictif en L1 et en L2, qu'ils comparent avec celui d'un ensemble de variables plus simples, à savoir la présence de certains verbes, la longueur des phrases et la proportion de certaines catégories de discours (ensemble 2).

Il ressort plusieurs leçons intéressantes de leurs expériences : (1) les variables syntaxiques (ensembles 1 et 2) se comportent nettement moins bien que les lexicales ; (2) les variables obtenues suite à une analyse syntaxique ont une capacité prédictive supérieure à celles calculées à partir d'une simple annotation en catégories ; (3) cependant, cette supériorité est

10. Schwarm et Ostendorf (2005, 526) utilisent pour cet acronyme la définition donnée dans le Penn Treebank, à savoir une proposition introduite par une conjonction de subordination, laquelle peut être omise.

limitée et, au vu de la lourdeur de l'analyse syntaxique, les auteurs concluent que les facteurs de l'ensemble 2 constituent une réelle alternative lorsque l'on ne peut se permettre un traitement automatisé lourd. Enfin, (4) il apparaît que les variables syntaxiques sont davantage corrélées avec le critère en L2 qu'en L1.

Nous pensons donc que ce type de variables pourrait se révéler particulièrement utile en FLE, pour autant que l'on dispose d'un analyseur syntaxique efficace pour le français et que le même travail de sélection de structures pertinentes soit effectué sur le FLE. À cet égard, on peut envisager deux solutions. D'une part, Heilman *et al.* (2008) proposent une technique de sélection automatisée, basée sur les arbres de dérivation d'un large corpus. Ils retiennent les 1 000 structures les plus fréquentes d'une profondeur variant de 0 à 3. D'autre part, les référentiels officiels du CECR actuellement en publication pourraient se révéler de précieux guides, puisqu'ils établissent des listes de structures cibles pour apprenants du FLE. Nous n'avons cependant pas retenu ces pistes, eu égard à leur temps de calcul exagéré par rapport à l'apport informationnel.

D. Les variables basées sur le modèle transformationnel : Dans son ouvrage *Syntactic Structures*, Chomsky (1957) postule l'existence d'un double niveau de représentation pour une phrase : une structure profonde et une structure de surface, qui est obtenue à la suite de transformations appliquées à la structure profonde. Selon ce modèle, un lecteur doit, pour comprendre un texte, effectuer l'opération inverse et détransformer chaque phrase d'un texte, retrouvant ainsi sa structure profonde. Plusieurs expériences rapportées par Schlesinger (1968) montrent que les phrases résultants de transformations seraient plus complexes à comprendre, à lire, à produire et à se remémorer que des constructions simples.

À partir de ce modèle, Bormuth (1969a) a imaginé se servir du nombre de transformations inverses nécessaires pour retrouver la ou les phrases simples correspondant à la structure profonde d'une phrase à lire. Il reprend les 50 constructions syntaxiques envisagées pour la famille de variable précédente et leur attribue un poids transformationnel. Dès lors, la difficulté d'un passage se calcule comme le nombre de transformations inverses nécessaires pour retomber sur les structures profondes de chaque phrase, divisées soit par le nombre de mots ($r = -0,423$), de propositions ($r = -0,604$), le nombre d'unités minimales de ponctuation ($r = -0,643$) ou encore par le nombre de phrases ($r = -0,640$). Dans les quatre cas, la corrélation avec le critère est significative.

Cette idée est envisagée pour le français par Henry (1975, 81-82), qui doute cependant que la notion de « poids transformationnel » soit pertinente. Il estime plutôt que les lieux de complexité sont les mots où l'hypothèse de lecture précédemment formulée par le lecteur se voit infirmée et qu'il convient de se focaliser sur les transformations qui entraînent un déplacement des constituants de leur position canonique.

Par la suite, ce type de variable n'a plus guère connu de succès, sans doute à cause du déclin du modèle transformationnel. En effet, Richaudeau (1981, 505), par exemple, estime que ses expériences sur la compréhension, mesurée à partir de la mémorisation, invalide les principes psychologiques de la grammaire générative. La difficulté d'une phrase ne serait pas corrélée avec la complexité de la structure en arbres ni le nombre de transformations requises pour former la phrase cible à partir de la phrase noyau, puisque l'emploi de « structures-avertissements » renforce la mémorisation. Or, dans la grammaire générative, ces structures « se traduisent graphiquement par des arborescences complexes » (Richaudeau, 1984, 95).

E. Les variables de poids structural : Cette cinquième famille de mesures repose sur les arbres de dérivation obtenus lorsqu'on analyse des phrases à l'aide d'une grammaire hors contexte. À nouveau, c'est Bormuth (1966) qui les envisage en premier, se reposant sur le concept de profondeur lexicale suggéré par Yngve (1960, cité par Schlesinger, 1968) pour la production et adapté à la réception par Miller et Chomsky (1963). Ce modèle postule un mécanisme linéaire de décodage dans lequel « la compréhension est donc conditionnée, à tout moment, par le poids de la partie du passage stockée en mémoire et par les contraintes syntagmatiques que la partie de la séquence déjà lue impose au reste de l'énoncé » (Henry, 1975, 84).

Plus concrètement, Bormuth (1966, 1969a) calcule, pour chaque mot, le nombre de noeuds conservés en mémoire à un moment donné (poids gauche), ainsi que les noeuds induisant une contrainte sur la suite de la phrase (poids à droite). Sur cette base, il définit 16 variables, telles que la somme des poids gauches ou droits normalisée en fonction du nombre de phrases ou de mots, etc. Toutes se révèlent significatives, leurs corrélations oscillant entre $-0,405$ et $-0,586$ (pour la somme des poids droits par phrase). Signalons tout de même que la corrélation du nombre de mots par phrase atteint $-0,605$ sur le même corpus, ce qui relativise l'intérêt des mesures de poids structural, en particulier vu les efforts nécessaires à leur paramétrisation. Bormuth (1966, 122) conclut pourtant à des apports informationnels en partie distincts.

Ce faible recouvrement informatif est mis en doute par l'étude de MacGinitie et Tretiak (1971), qui reprend cette famille de variables mais procède à des normalisations plus simples. Les auteurs se limitent à la profondeur maximale moyenne par phrase et à la somme des poids par phrase. Leurs résultats s'avèrent plutôt en défaveur des variables de poids structural, puisque non seulement leurs corrélations avec le critère sont inférieures à celle du nombre de mots par phrase, mais elles se révèlent également très corrélées à ce prédicteur classique.

Récemment, ce type de variables a toutefois retrouvé un certain succès. Graesser *et al.* (2004, 198) mesurent le nombre de constituants de niveau supérieur par mot. Il s'agit apparemment

du ratio de symboles non terminaux sur les symboles terminaux. Par ailleurs, Schwarm et Ostendorf (2005) et Pitler et Nenkova (2008) calculent la hauteur moyenne de l'arbre de dérivation. Les secondes rapportent toutefois une corrélation non significative pour ce facteur. Aussi, au vu des faibles performances de ce prédicteur, pourtant complexe à paramétriser, nous ne l'avons pas retenu pour notre étude.

F. La redondance des structures syntaxiques Enfin, de la même façon qu'il est pertinent d'évaluer la redondance des formes lexicales, il apparaît tout aussi intéressant de calculer la redondance des structures syntaxiques en lien avec la lisibilité. Kemper *et al.* (1993) adaptent à cet effet la mesure Index of Productive Syntax (IPsyn), présentée par Scarborough (1990) et qui visait à l'origine à estimer le développement des structures syntaxiques chez les jeunes enfants, c'est-à-dire dans un contexte de production. Cet indice évalue l'usage de 56 aspects syntaxiques, dont 27 formes de syntagmes nominaux et verbaux, 10 types de questions et de négations et 19 structures syntaxiques. Kemper *et al.* (1993, 413), qui l'emploient dans le cadre d'expérience en lisibilité avec des personnes âgées, ne sont cependant pas satisfaits de ses performances et renoncent à l'employer.

Plus récemment, Crossley *et al.* (2007) suggèrent de mesurer le taux de redondance des structures syntaxiques en lisibilité L2. En effet, ils estiment qu'une construction reconnue comme complexe, mais qui se répéterait plusieurs fois au sein d'un même texte, est moins susceptible d'entraîner des difficultés chez des apprenants. Ils paramétrisent cette variable sous la forme de la moyenne, pour chaque phrase adjacente du texte, de la proportion de noeuds en commun entre leurs arbres de dérivation. Ce facteur conduit à une corrélation nettement significative ($r = 0,71; n = 31; p < 0,001$) et est intégré au sein de leur formule. À nouveau, cette paramétrisation repose sur un analyseur syntaxique automatique, c'est pourquoi nous ne l'avons pas utilisée.

Les variables catégorielles

À côté des mesures de la longueur phrastique et de la complexité syntaxique, on trouve encore deux ensembles d'indices syntaxiques employés en lisibilité, dont le plus commun des deux est basé sur l'analyse morpho-syntaxique des textes, autrement dit sur les catégories grammaticales des mots qui composent un texte. L'hypothèse sur laquelle reposent ces variables est la suivante :

Hypothèse 3.6 *Tout texte consiste en une suite d'unités qui appartiennent à des classes et des sous-classes grammaticales particulières. La fréquence relative d'apparition de certaines classes et sous-classes serait indicatrice de la difficulté des textes.*

Nous avons trouvé deux preuves expérimentales en faveur de cette hypothèse. D'une part, Richaudeau (1984, 89) note, à partir d'expériences de mémorisation, que les mots les plus facilement oubliés ne sont pas les mots outils, mais les adjectifs et les adverbes. D'autre part, Hahne et Friederici (1999, 195) rapportent deux études basées sur les PEV, qui ont révélé la présence d'un composant gauche antérieur négatif lors du traitement des mots grammaticaux, lequel n'apparaît pas pour les mots appartenant à des classes ouvertes. Il semblerait donc que, non seulement, le traitement des mots diffère selon leur catégorie de discours, mais aussi que certaines d'entre elles soient mieux retenues que d'autres. Cela semble donc justifier l'emploi de variables catégorielles en lisibilité, même si leur effet précis sur la lecture reste méconnu.

Vogel et Washburne (1928) étaient apparemment du même avis, puisqu'ils intègrent le nombre de prépositions pour 1000 mots ($r = 0,518$) à leur formule finale. Par ailleurs, ils constatent que le nombre de verbes pour 1000 mots est aussi un bon prédicteur ($r = -0,527$), mais que c'est nettement moins le cas du nombre de conjonctions ($r = 0,296$). Notons que Washburne et Morphett (1938), lorsqu'ils révisent leur formule de 1928, en suppriment le nombre de prépositions. Ils ont en effet remarqué qu'il est étonnamment difficile pour la majorité des gens d'identifier l'ensemble des prépositions dans un texte.

À la même époque, Gray et Leary (1935) mentionnent des corrélations de $-0,261$ pour le pourcentage de mots lexicaux et de $0,261$ pour le pourcentage de mots grammaticaux. Ces résultats semblent indiquer que les textes complexes comprennent plus de mots-outils, mais moins de mots de contenus.

Le premier à envisager un grand nombre de catégories différentes est Bormuth. Il rapporte l'étude de Carroll (1960, cité par Bormuth, 1966, 90), lequel remarque que les catégories de discours mesurent une grande variété de caractéristiques stylistiques. Pour sa part, Bormuth (1966) expérimente deux ensembles de classes : les huit catégories classiques (nom, verbe, adverbe, adjectif, pronom, préposition, conjonction et interjection) et une classification à 5 classes proposée par Fries (1952, cité par Bormuth, 1966, 91). Sur la base de ces deux ensembles, il définit plus de 80 variables, dont les meilleurs se révèlent être le ratio de pronoms/conjonctions ($r = 0,805$), de verbes/conjonctions ($r = 0,727$) ou d'adverbes/conjonctions ($r = 0,663$). Il s'agit de facteurs assez inattendus, mais qui se révèlent nettement plus efficaces que ceux envisagés par les auteurs précédents : la proportion de verbes ($r = 0,421$), de noms ($r = 0,065$), de prépositions ($r = -0,421$) ou de conjonctions ($r = -0,441$).

Cette approche est ensuite reprise par quelques chercheurs en lisibilité du français [11]. Parmi ceux-ci, Henry (1975) calcule le pourcentage de prépositions simples sur la base d'une liste de 30 termes ainsi que le ratio de conjonction de coordination ; Daoust *et al.* (1996) proposent un décompte automatisé de 23 lexèmes ou locutions associés à des structures complexes. À notre connaissance, cette famille de prédicteurs n'a jamais été expérimentée dans le cadre du FLE, mais a été employée par Greenfield (2004) pour l'ALS.

Depuis l'apparition d'étiqueteurs morpho-syntaxiques automatisés (par ex. TreeTagger de Schmid (1994)), ces variables connaissent un plus grand succès. Ainsi, Graesser *et al.* (2004) identifient pas moins de 50 catégories différentes desquelles ils calculent le score d'incidence, à savoir leur proportion évaluée sur 1000 mots. Heilman *et al.* (2007) disent recourir à un nombre réduit de catégories, sans les spécifier, tandis que Feng *et al.* (2009) se limitent au nombre d'adjectifs par phrase et dans le texte. De leur côté, Pitler et Nenkova (2008) prennent en compte le nombre de pronoms et le nombre d'articles définis par phrase, mais relient ces deux variables à des aspects de cohésion. Aucune des deux ne s'avère significativement corrélée avec leur critère.

Pour conclure, soulignons que certains résultats apparaissent à première vue contradictoires. Chez Bormuth (1966), les verbes et les noms entretiennent une corrélation positive avec le critère, alors que Gray et Leary (1935) avaient observé une corrélation négative entre leur critère et les mots lexicaux. On retrouve la même inversion de signe entre les deux auteurs en ce qui concerne les mots-outils. Une explication possible de ce phénomène est que le mélange des différents mots à contenu conduise à une relation inversée entre leur nombre et la difficulté. Cette hypothèse est soutenue par la corrélation négative du ratio des adjectifs ($r = -0,430$) et la corrélation positive de la proportion d'adverbes ($r = 0,334$). Plus problématique est le cas des verbes, où le signe de la corrélation s'inverse entre Vogel et Washburne (1928) et Bormuth (1966). Cela nous montre qu'il faut interpréter ces résultats expérimentaux avec prudence et qu'ils ne devraient pas être utilisés pour en conclure des relations causales entre certaines catégories de discours et la difficulté d'un texte, quand bien même ils peuvent se révéler de bons indices. Néanmoins, ces corrélations, qu'elles soient positives ou négatives, apparaissent suffisantes pour justifier de répliquer ces facteurs, ce que nous avons fait dans cette étude.

11. Un résumé catégorie par catégorie est présenté par de Landsheere (1982, 316). Toutefois, il nous semble qu'il tombe dans le travers que nous dénonçons ci-dessous, à savoir conclure rapidement sur la relation entre une variable catégorielle et la difficulté sur la base de peu d'expériences.

Complexité des temps verbaux

Le dernier type de facteurs syntaxiques envisagés dans cette étude est la complexité des formes verbales. Il s'agit ici de considérer que l'emploi de certains temps ou modes verbaux rend les textes plus complexes à comprendre que d'autres. Bien qu'il soit difficile de déterminer exactement quels temps et quels modes sont plus susceptibles de poser des difficultés, des études ont en tout cas confirmé l'importance de cette variable en lecture.

Carreiras *et al.* (1997, 445) suggèrent que les temps et aspects verbaux jouent un rôle important dans la construction d'une représentation mentale d'un texte et, par conséquent, dans sa compréhension. Les marqueurs de temps et d'aspect apportent ainsi des informations essentielles permettant au lecteur de distinguer les éléments principaux et secondaires liés aux événements décrits par ces verbes. Une seconde fonction de l'aspect a été découverte par Truitt and Zwaan (1997, cités par Ferretti *et al.*, 2007) : il influence l'activation des connaissances générales portant sur des propriétés communément associées avec les événements décrits par les verbes. Ainsi, les auteurs observent que la lecture du mot « hammer » est davantage facilitée lorsqu'elle est précédée de la phrase « He was pounding the nail », où l'on trouve un verbe imperfectif, que pour la même phrase avec un verbe perfectif.

À notre connaissance, cet aspect de la difficulté syntaxique a peu été traité en lisibilité. Gunning (1952, 33) critique Gray and Leary en disant que ceux-ci ont ignoré des facteurs importants, dont les verbes. Toutefois, dans sa propre formule, il n'en tient pas davantage compte. Il dédie toutefois un chapitre entier de son ouvrage à cet aspect et propose parmi ses dix facteurs pour mieux écrire, le pourcentage de formes verbales fortes, c'est-à-dire, essentiellement, les verbes d'action (Gunning, 1952, 107-118). Par après, Gillie (1957) introduit le nombre de verbes finis dans sa formule du taux d'abstraction. Il faut ensuite attendre Daoust *et al.* (1996) pour retrouver ce facteur, cette fois paramétrisé comme le ratio des verbes conjugués sur le nombre de mots du texte. Les auteurs constatent que cette proportion est un indice de facilité et ils expliquent cela par la présence « d'un grand nombre de phrases courtes à construction simple du type sujet-verbe-complément » (Daoust *et al.*, 1996, 214).

En ce qui concerne la L2, Heilman *et al.* (2007) prennent en compte, pour l'anglais, les aspects verbaux suivants : le présent, le passé, le progressif, le perfectif et le continu. Cependant, ils ne rapportent pas de résultats spécifiques aux aspects verbaux, qui sont évalués au sein d'un ensemble de 12 variables comprenant la longueur phrastique. À notre connaissance, nous sommes le premier à avoir mis en évidence l'apport des temps et aspects verbaux au sein d'une formule de lisibilité pour le FLE (François, 2009a). Nous discuterons plus en détail nos résultats concernant ce point dans le Chapitre 5.

Conclusion

Dans cette section, nous avons mis en évidence deux caractéristiques des indices de lisibilité propres à la difficulté syntaxique. D'une part, depuis leur première utilisation par Vogel et Washburne (1928), qui sont d'ailleurs à l'origine de nombre d'entre eux, ce type de variables a connu un large succès. Bien peu de formules ont prétendu s'en passer et certaines ont même concentré leurs efforts sur cette dimension linguistique particulière (Botel et Granowsky, 1972; MacGinitie et Tretiak, 1971). C'est que, même s'ils restent en moyenne moins efficients que les prédicteurs lexicaux, les indices syntaxiques apportent une amélioration significative des performances lorsqu'ils sont combinés à des mesures lexicales au sein d'une formule de lisibilité. Selon les théories que nous avons présentées précédemment sur les processus de lecture en L2, cette efficacité devrait encore être accrue pour des apprenants de FLE. Toutefois, il ne semble pas y avoir d'étude qui ait traité de cette question expérimentalement.

D'autre part, nous avons observé une évolution dans le type de variables syntaxiques utilisées en lisibilité. À l'origine, on comptait manuellement des phénomènes tels que le type de phrases ou les catégories du discours. Puis, sous l'influence de la grammaire générative et transformationnelle, plusieurs chercheurs expérimentent des variables basées sur des modèles formels de la syntaxe. Toutefois, le pouvoir prédictif de ces nouvelles mesures n'est pas supérieur à celui de facteurs plus simples, alors qu'elles sont nettement plus complexes à dénombrer, autant en termes de temps qu'en termes des connaissances linguistiques requises. Cette complexité reste d'ailleurs un obstacle pour la lisibilité computationnelle, puisque les performances des analyseurs syntaxiques automatisés sont encore insuffisantes et ceux-ci fort gourmands en temps de traitement. Étant données ces diverses considérations théoriques et pratiques, nous avons moins expérimenté à ce niveau qu'en ce qui concerne les facteurs lexicaux.

3.2.4 Les variables de la difficulté conceptuelle

Si les dimensions lexicales et syntaxiques des textes ont fait l'objet de nombreuses études depuis les débuts de la lisibilité, le Chapitre 2 a montré que le paradigme structuro-cognitif avait stimulé l'exploration et la formalisation de nouveaux indices de la difficulté, parfois déjà ébauchés dans des travaux plus anciens. Les chercheurs issus de ce paradigme refusent de se limiter au lexique et à la structure. Ils veulent saisir la complexité engendrée par le contenu du texte et son organisation. Ce second aspect étant traité dans la prochaine section, nous nous attardons ici sur les mesures de la difficulté conceptuelle.

Évaluer la complexité du contenu d'un texte apparaît comme une tâche complexe. Il n'est

déjà pas évident de déterminer formellement le contenu d'un texte donné. L'humain appréhende évidemment celui-ci d'une façon implicite, mais il peine à formaliser les processus à l'oeuvre de telle manière qu'une analyse sémantique automatisée soit possible. Il existe bien sûr des systèmes d'analyse sémantique plus ou moins aboutis, mais leur développement reste en retrait en ce qui concerne le français et ils n'offrent pas des performances assez fiables pour un contexte pédagogique.

De plus, une fois définie une formalisation particulière du contenu d'un texte, il faut encore établir comment celle-ci entraîne des difficultés lors de la lecture. À nouveau, on se retrouve confronté à la dimension interactive du processus de lecture, qui importe tout particulièrement en ce qui concerne les aspects sémantiques. Or, rappelons que l'objectif des formules de lisibilité est de poser un diagnostic en amont de la réception des textes par leurs lecteurs. Dès lors, les mesures de la difficulté conceptuelle se sont concentrées sur des aspects textuels mesurables, et ce à deux niveaux différents : celui des mots et celui du treillis conceptuel qu'ils tissent en entrant en relation.

Dans la suite de cette section, nous allons tout d'abord envisager deux ensembles de mesures qui cherchent à évaluer la densité conceptuelle des textes en prenant en compte les caractéristiques des mots que ceux-ci contiennent, à savoir leur niveau de concrétude et leur polysémie potentielle. La question principale que cachent ces deux familles de variables est la façon dont nous accédons, à partir d'une forme orthographique, aux représentations sémantiques qui lui sont associées et dont nous sélectionnons la représentation adéquate en fonction du contexte. Elles ne disent rien de la façon dont les sens activés s'agencent pour former des idées plus ou moins complexes. Cet aspect est traité par une troisième famille de variables qui tentent de mesurer la densité des idées identifiables dans un texte. C'est avec celles-ci que nous clôturons cette section.

Le niveau d'abstraction du vocabulaire

On s'accorde aisément à dire qu'un texte comportant de nombreuses réflexions abstraites est en moyenne plus difficile à appréhender qu'un texte similaire dont le sujet est plus concret. Malheureusement, il n'est pas aisé de repérer les principales idées d'un texte ni de leur attribuer une valeur d'abstraction. C'est pourquoi, une fois de plus, c'est au travers du vocabulaire que les chercheurs ont principalement cherché à mesurer le niveau d'abstraction d'un texte. Notre hypothèse est donc la suivante :

Hypothèse 3.7 *Si les mots concrets sont lus plus vite que les mots abstraits, on peut supposer que c'est parce qu'ils réfèrent à des concepts plus accessibles pour le lecteur. Par conséquent, plus la proportion de mots concrets est élevée dans un texte, plus celui-ci sera*

aisé à comprendre.

Précisons que nous considérons ici les mots concrets comme « tous les mots référant à des personnes, des substances, des lieux, ou des choses pouvant être vues, touchées, senties, ou goûtées (Kammann et Streeter, 1971), tandis que les mots abstraits sont tous les mots ne pouvant pas être expérimentés par les sens » (Ferrand, 2007, 420).

Pour vérifier la validité de cette hypothèse, nous avons, une fois de plus, exploré la littérature en psycholinguistique. L'existence d'un effet de concrétude a été confirmée à plusieurs reprises (Bleasdale, 1987; Kounios et Holcomb, 1994; Schwanenflugel *et al.*, 1988) pour des mots isolés, mais Schwanenflugel *et al.* (1988, 499-500) rapportent aussi plusieurs expériences où le temps de lecture de phrases abstraites est supérieur à celui de phrases plus concrètes.

Plusieurs explications de ce phénomène ont été avancées et aujourd'hui encore la communauté scientifique n'est pas parvenue à un consensus. Les deux explications principales sont la théorie du double codage (Paivio, 1991) et la théorie de la disponibilité contextuelle (Schwanenflugel *et al.*, 1988). La première postule l'existence de deux systèmes de représentation : un verbal et un non verbal ou imaginal (basé sur les images). Ceux-ci sont fonctionnellement indépendants l'un de l'autre (c'est-à-dire que l'un peut être activé sans l'autre), mais leurs unités (logogen et imagen) sont interconnectées et peuvent s'activer réciproquement à l'apparition d'un stimulus approprié (Paivio, 1991, 258-259). Dans ce cadre théorique, l'effet de concrétude s'explique parce qu'un mot concret activerait les deux systèmes à la fois et atteindrait plus rapidement le niveau d'activation nécessaire, tandis que les mots abstraits ne sont pas associés à une image dans le système imaginal. Quant à la théorie de Schwanenflugel *et al.* (1988), elle fait l'hypothèse que l'effet de concrétude provient plutôt d'une différence qualitative au sein du système verbal, en ce sens que les mots concrets sont plus fortement associés à d'autres éléments du système. Dans ce cas, la disponibilité de l'information contextuelle est plus grande, ce qui facilite la reconnaissance des termes concrets.

Depuis lors, ces deux théories ont été éprouvées, notamment à l'aide des PEV et de techniques d'imagerie cérébrale, qui révèlent certains faits intéressants. Tout d'abord, étant donné que l'effet de concrétude est observé lors de la reconnaissance de mots, il est légitime de se demander s'il ressortit plutôt des représentations orthographiques, à l'instar des effets de fréquence ou de familiarité, ou de représentations sémantiques, comme le laissent supposer les deux hypothèses décrites ci-dessus. À ce sujet, Kounios et Holcomb (1994) remarquent que le composant N400, généralement associé aux processus sémantiques, est influencé par la distinction entre mots abstraits et concrets, puisque son amplitude est supérieure pour les mots concrets. Cela confirme l'hypothèse selon laquelle l'effet de concrétude

ressortit aux traitements sémantiques. Les auteurs estiment que cette différence d'amplitude indique une plus grande activation de l'information sémantique en mémoire dans le cas des mots concrets.

Deuxièmement, Kounios et Holcomb (1994) repèrent que l'avantage des mots concrets dans le N400 est plus large dans l'hémisphère droit que gauche. Ils interprètent ce résultat comme une preuve en faveur de l'hypothèse du double codage : les deux types de mots seraient traités par le système verbal, plutôt associé à l'hémisphère gauche, tandis que l'hémisphère droit correspondrait à l'activation du système imaginal, essentiellement utilisé pour les mots concrets (Kounios et Holcomb, 1994, 821). Binder *et al.* (2005) confirment ces deux hypothèses à l'aide d'une expérience qui recourt à l'IRMf dans une tâche de décision lexicale. Non seulement ils obtiennent un effet de concrétude significatif, qu'ils associent à un traitement sémantique, mais l'IRMf, plus précise que les PEV dans la localisation des régions activées par un stimulus, révèle effectivement une activation de l'hémisphère gauche pour les mots abstraits et une activation plus bilatérale pour les mots concrets.

Le niveau de concrétude d'un mot jouerait donc un rôle non négligeable en lecture L1 et il semblerait que ce soit aussi le cas en L2. En effet, De Groot (1992), entre autres, fait la preuve que les mots concrets sont traduits plus précisément et ceux de haute fréquence sont, de plus, traduits plus rapidement. Par ailleurs, la concrétude améliore non seulement le rappel immédiat chez les bilingues (Paivio *et al.*, 1988), mais facilite également l'acquisition du vocabulaire sous certaines conditions (Van Hell et Mahn, 1997). Enfin, van Hell et de Groot (1998) rapportent, pour des néerlandophones devant effectuer une tâche de décision lexicale en anglais, un effet de concrétude légèrement significatif, du moins lorsque l'effet de concrétude et la disponibilité contextuelle sont mêlés. Une fois ceux-ci distingués, l'avantage des mots concrets n'apparaît plus, ce qui conduit les auteurs à défendre la thèse de la théorie de la disponibilité contextuelle pour le cas bilingue.

Notons toutefois que Paivio *et al.* (1988) apportent de leur côté des preuves en faveur de la théorie du double codage. Celle-ci postule un système verbal pour chacune des deux langues, mais un système imaginal unique. Dès lors, comme ces trois systèmes de représentation sont interconnectés, les mots concrets présentent un avantage non négligeable : ils peuvent être traduits directement grâce à un lien entre les deux systèmes verbaux ou indirectement via le système imaginal (van Hell et de Groot, 1998, 42-43). Ce chemin additionnel expliquerait la plus grande facilité à traduire les mots concrets, ce qui est une étape nécessaire pour les lecteurs en L2.

Ainsi, si une explication unique de l'effet de concrétude n'est pas encore d'actualité, on voit que les chercheurs s'accordent du moins sur l'existence d'une différence de traitement. Cela justifie donc de prendre en compte ce facteur dans une formule de lisibilité, comme

cela a d'ailleurs déjà été fait à plusieurs reprises. La difficulté principale pour paramétriser cette variable est de distinguer les mots abstraits des mots concrets, ce qui ne constitue pas une tâche triviale. C'est pourquoi, à côté des expériences basées sur une liste de termes concrets/abstraits, on rencontre deux autres stratégies qui contournent cette problématique en employant soit des catégories de discours, soit les affixes comme indices de l'abstraction. Nous avons déjà traité de ces mesures dans les sections précédentes, mais nous les envisageons à nouveau ici comme des « proxys » du taux d'abstraction des textes. Enfin, les deux dernières catégories envisagées ci-dessous portent sur deux variantes du taux d'abstraction : le niveau d'imageabilité et le taux de personnalisation des textes.

A. Mesures basées sur des listes La technique la plus directe pour évaluer le niveau d'abstraction d'un texte consiste à calculer le pourcentage de mots abstraits qu'il contient. Toutefois, rappelons que la difficulté est alors de définir des critères explicites qui permettent de distinguer les termes abstraits des concrets. Ainsi, Gray et Leary (1935, 105-106) renoncent à employer ce facteur, car ils ne disposent pas de règles de classification suffisament objectives. À la même époque, McClusky (1934, 280) effectuent une tentative en ce sens, isolant les noms référant à des objets concrets : il en ressort que ceux-ci sont plus courants dans les textes narratifs que les textes informatifs. De son côté, Lewerenz (1935) calcule le pourcentage de mots « sensoriels », mais il ne fournit pas assez de précisions sur ses résultats pour qu'on puisse en conclure quoi que ce soit.

Les premières données concernant l'efficacité de ce prédicteur sont dues à Lorge (1939) qui évalue la pertinence des catégories III et IV de Morriss et Halverson (*cf.* Chapitre 2.2.2). La catégorie III comprend les mots acquis au cours de l'éducation, tels que « Iraq » ou « filament », c'est-à-dire des termes concrets, mais plus complexes que ceux repris dans les catégories I et II. Elle obtient une corrélation de $0,5526$ avec le critère, relativement similaire à celle de la catégorie IV, reprenant les mots abstraits ($r = 0,5739$). Signalons que ces résultats sont parmi les meilleurs obtenus par Lorge (1939), puisque seuls le nombre de mots par phrase ($r = 0,5988$) et le pourcentage de mots difficiles ($r = 0,5958$) les surpassent.

Ces résultats intéressants sont confirmés par trois autres études anglophones et une francophone. D'abord, Crane (1963, cité par Klare, 1974, 71) entreprend de remplacer le nombre d'absents de la liste de 3000 mots de Dale par d'autres facteurs, parmi lesquels il envisage la quantité de verbes abstraits. Il rapporte un R de $0,924$ pour sa formule à trois variables, par rapport à des jugements humains sur 9 articles, mais sans préciser l'effet des verbes abstraits considérés isolément. Coleman (1968, cité par Henry, 1975, 78) est plus précis et rapporte une corrélation de $-0,73$ entre le nombre de noms abstraits et son critère. Il défend par ailleurs l'efficacité de ce prédicteur, disant que « l'effet des noms abstraits est assez impor-

tant pour contrebalancer le manque de fidélité dû à des problèmes de définition » (Henry, 1975, 78). De même, Elley (1969) observe, pour cette même variable, une corrélation médiane de $0,81$ sur cinq échantillons de textes différents. Soulignons que comparativement à ses autres variables, il s'agit cependant d'un résultat moyen.

Pour le français, Henry (1975) se base sur les critères définis par Gorman (1961) pour distinguer les mots abstraits et concrets. Sur son corpus, le pourcentage de noms concrets se comporte très favorablement, atteignant des corrélations de $0,66$, $0,6$ et $0,52$ pour les trois niveaux scolaires. Ces résultats indiquent par ailleurs que l'importance de la concrétude semble diminuer avec l'âge. Malheureusement, la technique de paramétrisation utilisée par Henry repose sur une analyse linguistique dépendante du contexte et est difficilement reproductible. Aussi en a-t-il établi une seconde qui recourt à une liste de noms concrets. Celle-ci a été obtenue à partir de la liste de 7918 noms de Ters, Mayer et Reichenbach (1964, cités par Henry, 1979, 56) et conduit à des performances relativement comparables à la technique précédente (le r entre les deux est de $0,85$), tout en étant automatisable.

Enfin, plus récemment, Graesser *et al.* (2004) se reposent sur la base de Coltheart (1981, cité par Graesser *et al.*, 2004) : la MRC Psycholinguistics Database, qui contient plus de 150 000 mots et fournit diverses informations lexicales, dont le niveau de concrétude, d'imageabilité et de saillance de l'information. Ils ne rapportent toutefois pas de corrélations entre ces trois facteurs et une variable dépendante.

En résumé, les mesures basées sur des listes de mots abstraits ou concrets semblent constituer de bons prédicteurs pour la lisibilité des L1, mais nous n'avons pas réussi à mettre la main sur une étude expérimentale dans le contexte L2, ni sur une liste suffisamment complète pour le français.

B. Mesures basées sur des catégories de discours Lorsque nous avons envisagé la question des catégories de discours, nous avons insisté sur la difficulté de déterminer quels aspects précis du processus de lecture sont mesurés par ce type de facteurs. Comme nous l'avons détaillé à la section 2.2.3, Flesch (1950) a émis l'hypothèse que le niveau d'abstraction d'un texte transparaissait dans la répartition des catégories de discours. Il est suivi sur ce point par Gillie (1957), qui simplifie les 16 règles de comptage de Flesch pour ne retenir que le nombre de verbes finis et d'articles définis.

Cette hypothèse inspire Bloomer (1959), qui suppose que les adjectifs et adverbes serviraient à préciser le sens des mots vagues et que leur nombre augmenterait proportionnellement au niveau d'abstraction. Que cette hypothèse soit correcte ou non, ce chercheur obtient bien une corrélation de $-0,61$ avec son critère pour le ratio de mots par modifieur. Il cite aussi un article de Broder (1940, cité par Bloomer, 1959, 269) qui a détecté que le ratio adjectifs/verbes

augmente également en fonction du taux d'abstraction.

Mentionnons enfin les travaux de Hull (1979, cité par DuBay, 2004, 51), lequel postule qu'un plus grand nombre d'adjectifs et d'adverbes situés avant un nom diminue le taux de compréhension, et relève bien un effet négatif ($r = -0,644$) de la charge de ces modifieurs sur celle-ci. Par la suite, nous n'avons plus connaissance d'études approchant le niveau d'abstraction d'un texte via les catégories de discours, ni en L1, ni a fortiori, en L2.

C. Mesures basées sur le nombre d'affixes Nous avons vu précédemment (*cf.* Section 3.2.2) que les affixes avaient été employés par Flesch (1943, cité par Chall, 1958) comme un indice de la complexité lexicale. Toutefois, Henry (1975, 77) estime qu'un « certain nombre de suffixation nominales sont caractéristiques de noms à caractère abstrait plus ou moins accusé ». Il liste 25 suffixes qui lui servent à définir un pourcentage de mots suffixés qu'il ne retient toutefois dans aucune de ses formules. Cette technique avait déjà été employée par Gillie (1957, 216), en vue de simplifier l'indice d'abstraction de Flesch (1950). Lui aussi détaille sept suffixes à partir desquels il détermine une estimation du nombre de noms abstraits. Gillie ne calcule cependant pas de corrélation avec un critère ; seulement avec l'indice d'abstraction de Flesch. Celle-ci est d'ailleurs relativement élevée : $r = -0,7411$. De manière générale, ce facteur n'a toutefois pas connu un grand succès et nous ne l'avons pas conservé pour la suite de notre étude.

D. L'imageabilité Un aspect lexical très proche de la concrétude est l'imageabilité, c'est-à-dire la facilité à se former mentalement une image d'un mot (Ferrand, 2007, 420). Les deux dimensions seraient apparemment très corrélées entre elles : Schwanenflugel *et al.* (1988) rapportent dans leurs expériences des corrélations supérieures à $0,90$ entre ces deux facteurs. D'ailleurs, certains auteurs utilisent les deux termes comme des synonymes et Lupker (2005, 54) dit douter de l'intérêt de la distinction qui lui apparaît quelque peu artificielle.

Ce qui semble en tout cas certain, c'est que les effets de l'imageabilité sur la lecture sont similaires à ceux de la concrétude. Citons, entre autres, l'article de Steinel *et al.* (2007) sur les effets de l'imageabilité en L2, lesquels concluent à une amélioration significative des performances lors de l'apprentissage de paires bilingues où les mots sont associés. Un résultat secondaire de cette étude, particulièrement intéressant pour notre problématique, est la sensibilité supérieure des sujets à l'imageabilité lors d'un apprentissage réceptif que productif. Autrement dit, cet aspect jouerait plus à la réception, ce qui est le cas lors de la lecture en L2.

Sans surprise, les auteurs se réfèrent à la théorie du double codage pour expliquer les effets de cette dimension lexicale, ce qui renforce l'idée que la différence avec la concrétude est mince. En lisibilité, les seuls à avoir envisagé ce facteur sont Graesser *et al.* (2004, 196),

toujours sur base de la *MRC Psycholinguistics Database*. Comme nous l'avons déjà précisé, ils se limitent cependant à l'étape de paramétrisation.

E. Taux de personnalisation du texte Un dernier aspect qui nous semble pouvoir être rapproché de la concrétude est le taux de personnalisation des textes. Cette dimension linguistique a connu plusieurs approches depuis que Dale et Tyler (1934, 395) ont émis l'hypothèse que les textes rédigés dans un style plus informel sont plus simples à lire. Ces auteurs ont mesuré le niveau de formalité des textes via le nombre de pronoms personnels et ils obtiennent des résultats mitigés : ceux de la 2e personne semblent en effet rendre plus aisée la lecture ($r = 0,251$), tandis que ceux de la 1re personne ($r = -0,091$) et de la 3e personne ($r = -0,069$) présentent une corrélation négative avec le critère, probablement non significative.

Cette première expérience soulève une question importante de paramétrisation : est-il justifié de fusionner certaines classes de pronoms ? En effet, pour le français, Henry (1975, 85) rassemble pronoms de la 1re et de la 2e personnes et, alors que les effets devraient s'annuler d'après Dale et Tyler (1934, 395), la variable est suffisamment informative pour que le chercheur belge l'intègre dans 6 formules sur 9. De même, Gray et Leary (1935) testent à la fois les trois classes séparément et conjointement pour l'anglais. Ils rapportent des corrélations positives dans les différents cas, ce qui confirme la possibilité de fusion. Par contre, Daoust *et al.* (1996) s'aperçoivent que si le « tu », lié à un univers plus familier, est bien un indice de facilité des textes, c'est l'inverse pour le « vous », qui annonce une forme conjuguée plus complexe. En conséquent, il conviendrait de les distinguer. Précédemment, nous avons toutefois repris la paramétrisation d'Henry, sans obtenir un prédicteur très efficace (François, 2009a). Nous avons dès lors décidé d'essayer, dans cette étude, les différentes classes possibles.

Notons que cette approche limitée du taux de personnalisation a été étendue par Flesch (1948, 229), lequel ne considère pas seulement les pronoms personnels au sein de sa formule d'intérêt personnel, mais également les noms de genre naturel féminin ou masculin (ex. Jones, Mary, père, actrice, etc.). Dale et Chall (1948a, 14) ont cependant objecté que cette paramétrisation présente le même désagrément que les noms propres dans l'effet de fréquence : selon qu'ils soient intratextuels ou fassent référence au réel, leur effet sur la lecture varie probablement.

Conclusion Qu'il soit mesuré à l'aide de listes, des affixes ou des catégories de discours, le niveau de concrétude des mots apparaît bel et bien comme un prédicteur efficace dans le contexte L1. Nous avons ainsi décrit plusieurs expériences qui, bien qu'elles emploient des textes et des critères divers, permettent d'observer des corrélations assez élevées pour chacune de ces trois techniques. Par contre, les indices basés sur le taux de personnalisation du

texte ne semblent guère discriminants dans les quelques études qui les envisagent, d'autant que celles-ci ne s'accordent pas sur la direction de l'effet des différents pronoms.

Par contre, on regrettera qu'aucune étude n'ait jamais porté sur la capacité de discrimination de ces indices dans un contexte L2. D'autant que les considérations théoriques que nous avons relevées dans la première partie de cette section laissent penser qu'il s'agirait d'une variable intéressante. Nous avons dès lors cherché à obtenir une liste distinguant les mots concrets et abstraits. Malheureusement, les listes disponibles souffrent d'une couverture trop faible ou sont incomplètes. Ainsi, la liste de Desrochers et Bergeron, 2000 ne détaille les valeurs d'imagerie que pour 1916 mots, tandis que le DELAF (Coste, 1990) ne donne les valeurs de concrétude que pour les noms composés. Par ailleurs, nous avons envisagé une procédure d'acquisition automatique des noms abstraits à partir d'un dictionnaire informatisé (Kister et Jacquey, 2007), mais celle-ci présente des performances trop faibles (62% d'étiquetages corrects) pour être utilisée à fin de constituer une ressource pour la lisibilité. Nous avons donc dû renoncer pour l'instant à paramétriser ce prédicteur, qui mériterait cependant de l'être.

Effets du nombre de sens d'un mot

Divers chercheurs se sont penchés sur un autre facteur qui influence le traitement sémantique : la polysémie lexicale. Autrement dit, est-ce que les mots ambigus posent davantage de problèmes que ceux qui ne possèdent que peu de sens différents ? Il semblerait que la polysémie joue un rôle différent, selon que la tâche à accomplir mobilise uniquement les représentations orthographiques (ex. : prononciation, décision lexicale) ou également les représentations sémantiques (compréhension de phrase à la lecture ou analyse du mouvement des yeux). Dans le premier cas, plus un mot comporte de sens, plus un lecteur le reconnaîtra rapidement (Millis et Button, 1989), mais lorsqu'il est nécessaire de sélectionner un sens en fonction du contexte, la pluralité sémantique augmente les temps de réaction (Rayner et Duffy, 1986). Nous allons discuter plus en détail ces deux effets et les explications qui en sont proposées.

Avant tout, il faut considérer la paramétrisation de cette variable. En effet, la technique qui vient directement à l'esprit est de consulter un dictionnaire et de compter le nombre de sens listés sous chaque entrée. Cependant, Gernsbacher (1984, 274-275) a montré que cette métrique de la polysémie ne convient pas. Non seulement l'auteur n'obtient pas d'effet notable pour la polysémie en employant cette approche, mais elle souligne que sa réalité psychologique est improbable. Qui parmi nous peut se targuer de connaître tous les sens des diverses formes lexicales qu'il a mémorisées ? D'autant que d'un dictionnaire à l'autre, le nombre de sens diffère.

C'est pourquoi Millis et Button (1989) ont exploré d'autres formes de paramétrisation pour ce facteur. Ils suggèrent les trois techniques suivantes :

First-meaning metric Les participants doivent écrire le premier sens qui leur vient à l'esprit et on calcule le nombre total de sens différents pour chaque mot.

Total-meaning metric Les sujets disposent d'un temps illimité pour lister l'ensemble des sens qu'ils connaissent pour chaque mot et on retient le nombre total de sens pour l'ensemble des informateurs.

Mean-meaning metric Dans les mêmes conditions que précédemment, la métrique est cette fois la moyenne du nombre de sens par sujet.

Les auteurs obtiennent des effets significatifs pour la seconde et la troisième métrique, ce qui confirme à la fois l'effet de polysémie lors d'une tâche de décision lexicale et les conclusions de Gernsbacher (1984) concernant l'inadéquation d'un dictionnaire comme référence du nombre de sens par mots.

Ces résultats ont été reproduits par d'autres chercheurs, dans des tâches de décision lexicale (Borowsky et Masson, 1996; Hino et Lupker, 1996; Hino *et al.*, 2002) et, dans une moindre mesure, à la prononciation (Hino *et al.*, 2002) [12]. Plusieurs explications à cet effet d'ambiguïté ont été apportées. Borowsky et Masson (1996) ont déterminé que l'effet d'ambiguïté se situait plutôt au niveau sémantique qu'orthographique et ils en simulent divers résultats à l'aide d'un modèle connexionniste distribué, proche de celui de Seidenberg et McClelland (1989), mais qui comporte un module sémantique en plus. Hino et Lupker (1996) expliquent la présence de cet effet facilitateur par une rétroaction du niveau sémantique vers le niveau orthographique. Un mot ambigu possède une représentation orthographique unique reliée à plusieurs représentations sémantiques. Dès lors, quand ces représentations sont activées, la quantité de rétroaction d'activation vers le niveau orthographique est supérieure à un mot monosémique et sa reconnaissance est plus rapide (Lupker, 2005, 54).

Cependant, comme nous l'avons expliqué précédemment, si la polysémie facilite la reconnaissance hors contexte, cette caractéristique apparaît inhibitrice en contexte, comme le révèlent les études effectuées via une technique d'enregistrement des yeux et de compréhension de phrases (Borowsky et Masson, 1996; Rayner et Duffy, 1986; Rayner et Frazier, 1989) ou via une tâche de catégorisation sémantique (Hino *et al.*, 2002). Cet effet inhibiteur serait lié à la manière dont on accède aux différents sens d'une même forme, laquelle serait fonction du contexte où la forme est rencontrée, ainsi que de la fréquence de ses différents sens (Duffy *et al.* 1988, cités par Degani et Tokowicz, 2010, 1267) [13]. Dans leur *modèle de la recherche réordonnée*, ces auteurs expliquent que si le temps de traitement sémantique

12. Notons que Borowsky et Masson (1996), par exemple, n'observent pas d'effet pour cette tâche.

13. D'autres théories ont également été avancées pour expliquer la procédure d'accès au sens (cf. Tabossi et Zardon, 1993), mais elles sont antérieures et nous sont apparues moins pertinentes.

des mots polysémiques est plus long, c'est dû à la compétition qui découle de l'activation parallèle de plusieurs sens, ce qui peut arriver dans deux cas : soit lorsque les deux (ou plus) sens en compétition sont équiprobables, soit lorsque le sens le plus fréquent apparaît dans un contexte qui biaise l'interprétation en faveur du sens moins fréquent.

Ce tableau, déjà complexe, est encore rendu plus ardu à dépeindre lorsque l'on considère le cas de l'apprenant en L2. En effet, dans le cadre d'un accès non-sélectif [14] aux différentes langues connues par l'apprenant, des phénomènes d'ambiguïté interlangues peuvent apparaître (Degani et Tokowicz, 2010). C'est le cas lorsqu'une même forme est partagée entre les deux langues, mais possède un sens différent dans chacune d'elles (faux-amis). Dans ce cas, Degani et Tokowicz (2010) suggèrent qu'une compétition entre les sens dans les deux langues va s'engager et que le choix du sens adéquat s'effectue non seulement en fonction de la fréquence des sens et du contexte sémantique, mais également de l'appartenance linguistique du contexte, autrement dit de la langue qui sert de support au mot ambigü.

Pour intéressantes qu'apparaissent ces observations sur la lecture des termes polysémiques en L1 et en L2, nous voyons également que paramétriser ce facteur dans le cadre de la lisibilité semble ardu. Si l'on voulait prendre en compte les effets interlangues, il conviendrait de constituer un modèle pour chaque paire de langues considérées, ce qui dépasse les objectifs de notre thèse. Une seconde option, plus simple et plus polyvalente, consiste à appliquer une perspective L1 telle que décrite ci-dessus à la L2. Il s'agit toutefois d'une hypothèse peu valide psychologiquement parlant. Elle a pourtant été adoptée par Graesser *et al.* (2004, 198), auteurs de la seule paramétrisation de ce problème en lisibilité. Ils ont par ailleurs effectué une seconde simplification, à savoir reprendre l'ensemble des sens d'un dictionnaire, et plus précisément WordNet (Fellbaum, 1998).

Comme Graesser *et al.* (2004) ne rapportent pas d'expérimentations pour ce facteur, il n'est pas possible de se faire une idée de son efficacité lorsqu'on le paramétrise de cette manière. Bien qu'il s'agisse d'une perspective à explorer, nous n'avons pas retenu ce facteur pour notre étude, car il est apparu comme trop difficile à paramétriser de manière conforme aux connaissances dans le domaine.

Mesures de la densité conceptuelle

Au-delà de la difficulté véhiculée par un mot rare ou un concept abstrait, il arrive que l'incompréhension naisse de l'association des termes – pourtant compris indépendamment – qui forment le treillis textuel. Chall et Dale (1995, 68-69) reconnaissent d'ailleurs que leur formule ne se comporte pas correctement lorsqu'un texte exprime des idées complexes à

14. Ce cadre théorique, dont nous avons déjà parlé précédemment, semble le plus accepté aujourd'hui.

l'aide de mots simples. Il s'agit certes d'une situation particulière, mais qui met en évidence un aspect qui a échappé à la majorité des formules de lisibilité : la densité des idées dans un texte.

Le premier chercheur qui a été sensible à cette question est McClusky (1934), à travers sa variable « nombre d'idées pour 100 mots ». Il pose déjà la question du repérage des idées, un processus dont il reconnaît la subjectivité. Il effectue celui-ci manuellement à la surface même des textes de son corpus (autrement dit, il ne postule pas de structure propositionnelle à la manière de Kintsch et van Dijk, 1978). Nous reproduisons ici un exemple tiré de son article (McClusky, 1934, 278) et reprenons sa notation, qui sépare chaque idée à l'aide du signe « / » :

> *Europeans often ask / how it happens that the Presidency of the United States / is not more frequently filled / by great and striking men.*

Au terme de cette première expérience, McClusky (1934) n'observe pas de réelles différences entre les textes simples et complexes. Toutefois, il est probable que cette absence de résultats provienne soit de la technique de paramétrisation, soit du corpus utilisé qui ne comporte que six textes.

Cette variable a ensuite inspiré Dolch (1939), qui s'attelle à repérer les « faits » dans des manuels scolaires. Il considère les « faits » comme les idées ressortissant à la discipline ciblée par le manuel. Il ne s'intéresse pas à l'ensemble des idées, car il estime que la difficulté naît plutôt des « faits ». Pour repérer ceux-ci dans les textes, il utilise le critère suivant : considérer comme un « fait » un contenu qui pourrait faire l'objet d'une évaluation. Dès lors, il compare leur densité au sein de manuels de géographie, d'histoire, de sciences et de santé et observe une densité décroissante, qu'il explique par un aspect narratif plus important dans les manuels de sciences et, surtout, de santé. Cependant, l'étude reste exploratoire et ne propose pas de conclusions en ce qui concerne le lien entre le niveau des manuels et la densité des « faits ».

Il faut attendre les études de Kintsch pour que soit définie une mesure plus systématique et reproductible de la densité conceptuelle. Celle-ci repose sur l'analyse prédicative de Kintch (1972, cité par Kintsch et Keenan, 1973) que nous avons décrite au Chapitre 2.2.6. À partir de cette théorie, Kintsch et Keenan (1973) montrent que la densité de propositions dans une phrase est une variable indépendante utile pour prédire le temps de lecture et la qualité du rappel. Sur des phrases dont la taille est contrôlée (de 14 à 16 mots), le temps de lecture augmente linéairement en fonction du nombre de propositions.

Au-delà de son intérêt vis-à-vis de la question de la densité, ce résultat renforce également l'hypothèse que la longueur des phrases n'est pas une cause de la difficulté (*cf.* Section 3.2.3)

et que ce serait plutôt la quantité de propositions qui importerait. En effet, la seconde expérience de Kintsch et Keenan (1973), où sont employées des phrases de longueur différente, confirme qu'au plus il y a de propositions dans une phrase, plus celle-ci demande de temps pour être lue et retenue. Les auteurs remarquent aussi que le temps nécessaire par proposition augmente pour les phrases plus longues : d'environ $1,3$ secondes pour les phrases de 10 mots, ce temps passe à $4,3$ secondes en moyenne pour des fragments textuels de 43 à 58 mots. Cette variation s'expliquerait parce que la taille de la structure propositionnelle conservée en mémoire augmente [15].

Dans une seconde étude, Kintsch *et al.* (1975) répliquent et approfondissent les résultats de Kintsch et Keenan (1973). Ils démontrent qu'une fois le nombre de propositions contrôlé, le nombre d'arguments différents dans une phrase influencent significativement le temps de lecture et, dans une moindre mesure, la qualité du rappel. Ainsi, les paragraphes où les mêmes concepts sont répétés sont plus aisés à lire que ceux où de nouvelles idées sont introduites sans cesse. On rejoint ici, d'une certaine manière, le type-token ratio, même s'il faut bien distinguer les deux mesures. Tandis que le TTR vise à estimer l'effet de facilité qui naît de la répétition des mêmes **formes lexicales** (y compris les mots grammaticaux), la densité des concepts s'applique à la facilité issue de la répétition des **concepts**, lesquels se réalisent ou non au moyen d'une même forme. Kintsch *et al.* (1975, 204) montrent que ces deux types de répétitions – de surface et conceptuelle – paraissent améliorer la rétention, mais précisent que la nature de leurs expérimentations ne permettent pas d'être assertif sur ce point.

Sur la base de ces deux études, Kintsch et Vipond (1979) définissent, au sein de leur formule de lisibilité (*cf.* Chapitre 2.2.6), deux prédicteurs qui mesurent la densité des concepts par mot et leur redondance. Il s'agit du nombre de propositions pour 100 mots et du nombre moyen d'arguments par proposition. D'autres auteurs de l'époque sont également sensibles à la densité conceptuelle : dans leur « checklist » pour la lisibilité des textes, Irwin et Davis (1980) insistent sur la nécessité d'introduire les concepts un à un, en les illustrant suffisamment à l'aide d'exemples. Ils ne proposent toutefois aucune mesure concrète. De la même manière, Amiran et Jones (1982) voient la lisibilité comme un espace à trois dimensions, dont l'une d'elles est la densité conceptuelle. Ils ne fournissent pas davantage une technique de paramétrisation pour la densité conceptuelle, suggérant simplement que les textes autoréférentiels sont plus complexes. Par « autoréférentiel », ils entendent un texte qui décrit une série de concepts, ensuite reliés entre eux (ex. un manuel de chimie où l'on rencontre des phrases du type : « la réaction *x* est similaire à la réaction *m*, si ce n'est que la molécule *a* y

15. On notera que ce résultat entre en contradiction avec les études précitées qui doutent que la phrase soit l'unité de traitement utilisée par l'humain. Dans ce cas, Kintsch et Keenan (1973) constatent bien un ralentissement progressif à mesure que la phrase s'allonge, laissant supposer que certaines informations sont mémorisées sur la phrase entière.

est substituée par la molécule *b* ») .

Il faut attendre Kemper *et al.* (1993) pour qu'ait lieu la première expérimentation en lisibilité à partir de ce cadre théorique. Les auteurs recourrent à une procédure décrite par Turner et Greene (cités par Kemper *et al.*, 1993, 410) pour décomposer les textes de leur corpus de test en une liste de propositions. Ils utilisent ensuite comme variable le nombre de propositions pour 100 mots. Celle-ci est corrélée à $0,52$ avec la formule de Dale et Chall (1948a), mais seulement à $-0,27$ avec la mesure de compréhension employée par les auteurs. Ils effectuent aussi une corrélation de premier ordre en contrôlant les connaissances en vocabulaire. Dès lors, la relation entre la densité propositionnelle et la compréhension devient non significative ($r = 0,07$), ce qui suggère que la compréhension de textes propositionnellement denses est liée aux connaissances lexicales. Cette étude semble dès lors assez mitigée en ce qui concerne l'efficacité de la densité propositionnelle comme variable prédictive de la lisibilité.

Cela explique peut-être pourquoi on ne trouve aujourd'hui guère de travaux qui s'intéressent à la densité conceptuelle, tandis que les variables de cohésion sont très utilisées. Relevons toutefois les travaux de Feng *et al.* (2009, 2010), qui avancent le concept de « densité des entités ». Ils postulent ainsi que le nombre d'entités introduites dans un texte est lié à la charge en mémoire de travail et influence dès lors la compréhension. L'avantage de cette vision est qu'elle est plus aisément automatisable. En effet, ils détectent automatiquement les entités nommées de personnes, d'organisation et de lieu, ainsi que les noms communs à l'aide d'un taggeur.

Au départ des entités, Feng *et al.* (2009) définissent les quatre variables suivantes : le nombre moyen d'entités par phrase et dans le texte, le nombre moyen d'entités uniques (type) par phrase et dans le texte. À partir de textes simples et complexes, les auteurs effectuent un test T de Student [16] qui se révèle significatif pour chacune des variables. Feng *et al.* (2010) explorent quant à eux 12 variables de densité, dont le pourcentage d'entités par phrases et par document ; la proportion d'entités nommées parmi l'ensemble des entités, etc. Il est intéressant de noter que, paramétrisées de cette manière, les variables de densité sont les plus prédictives parmi les facteurs cognitifs considérés dans l'étude. Combinées au sein d'une SVM (*cf.* 5), elles parviennent à une exactitude de 59% pour un problème à 4 classes.

Face à ces derniers résultats encourageants, nous avons décidé de tenir compte de la densité conceptuelle dans notre formule de lisibilité en L2. À notre connaissance, cet aspect de la difficulté conceptuelle n'a pas jamais été envisagée dans ce contexte. Nous utiliserons pour ce faire le logiciel *Densidées* (Lee *et al.*, 2010), qui est décrit à la Section 3.3.3.

16. Il s'agit d'un test de paramétrique qui compare les moyennes de deux échantillons afin de déterminer s'il est possible qu'ils proviennent de la même population ou non. Pour plus de détails, consulter Howell (2008) .

3.2.5 Variables de la cohérence textuelle

Après la complexité des concepts et leur densité, il faut encore envisager une autre source de difficulté sémantique : celle qui provient de la manière dont les concepts sont agencés au sein du texte. En effet, un écrit dont les relations logiques sont clairement explicitées et bien organisées est généralement plus aisé à appréhender qu'une suite de phrases faiblement connectées. Cet aspect, qui permet de distinguer un texte d'une suite de phrases sans liens entre elles, a été d'abord été étudié localement (entre phrases) et repris par Halliday et Hasan (1976) sous la notion de cohésion. Par la suite, la cohésion a été remise en cause et articulée au sein du champ plus vaste de la cohérence, qui comprend aussi l'organisation macrostructurelle des textes. Dans la suite de cette section, nous détaillons ces deux notions théoriques, avant de discuter leur importance lors de la lecture. Nous terminons en rapportant les travaux en lisibilité qui ont cherché à les paramétriser d'une manière ou d'une autre.

Cohérence et cohésion

Dans les années 70 et 80, plusieurs linguistes se sont focalisés sur un niveau supérieur à celui, traditionnellement étudié, de la phrase. On rassemble parfois leurs travaux sous le terme de « grammaires des textes ». Ceux-ci comprennent entre autres l'analyse propositionnelle de Kintsch (Kintsch et Keenan, 1973; Kintsch et van Dijk, 1978), les travaux de Van Dijk (1980) sur la macrostructure, la « story grammar » de Rumelhart (1975, cité par Carrell, 1982, 480), et surtout la théorie de la cohésion de Halliday et Hasan (1976).

Dans leur ouvrage, Halliday et Hasan (1976, 4) définissent ainsi la cohésion :

> *The concept of cohesion is a semantic one ; it refers to relations of meaning that exist within the text, and that define it as a text. Cohesion occurs where the INTERPRETATION of some element in the discourse is dependent on that of another. The one PRESUPPOSES the other, in the sense that it cannot be effectively decoded except by recourse to it.*

Par exemple, dans les deux phrases suivantes de Halliday et Hasan (1976, 2), traduites par nos soins :

> *Laver et peler six pommes à cuire. Les mettre dans un plat à four.*

le pronom « Les » réfère par anaphore au syntagme « six pommes à cuire » et n'est pas interprétable sans celui-ci. C'est la relation entre ces deux éléments qui fait de ces deux phrases un ensemble cohérent, c'est-à-dire un texte. Ces paires d'éléments entretenant une relation de cohésion sont appelées des « noeuds » (*ties*) et sont classés en cinq catégories :

les références (les référents des deux éléments sont reliés), les substitutions, les ellipses, les conjonctions (une relation sémantique est explicitée) et la cohésion lexicale (deux éléments partagent un même champ lexical).

Cette vision de la cohésion a cependant été assez rapidement battue en brèche (Carrell, 1982; Morgan et Sellner, 1980) à un point tel que Sanders et Maat (2006, 592) écrivent : « il est désormais clair que l'approche cohésive des liens textuels est inadéquate ». Parmi les premiers détracteurs de cette théorie, Morgan et Sellner (1980) considèrent que la présence de noeuds cohésifs n'est pas à l'origine de la cohérence textuelle, mais plutôt qu'elle la reflète. Ainsi, si l'on reprend nos deux phrases traitant de cuisine, on peut se demander comment un lecteur associe « les » à la suite de mots « six pommes à cuire », puisque « les » peut renvoyer à de nombreuses autres choses. Pour Morgan et Sellner (1980, 180), c'est parce que le lecteur fait l'hypothèse que le texte est cohérent qu'il est alors capable d'établir ce lien, en s'appuyant sur ses connaissances du monde réel et notamment sur les schémas (*cf.* Chapitre 1) qu'il a intégrés.

La cohérence postule donc que les connexions du discours ne se situent pas au niveau du texte, mais sont établies au sein des représentations mentales des lecteurs (Sanders et Maat, 2006, 592). Cette vision ne rejette cependant pas complètement des apports de Halliday et Hasan (1976, 4), car si la cohérence est plutôt de nature cognitive, sa construction mentale s'appuie aussi en partie sur les indices linguistiques présents dans le texte. Ceux-ci relèvent de deux niveaux distincts : la cohérence référentielle (qui régit les syntagmes) et la cohérence relationnelle ou relation rhétorique (qui régit des propositions) (Sanders et Maat, 2006, 593). Cette distinction correspond à celle qu'Armbruster (1984) opère entre les cohérences locale et globale, termes que nous avons repris dans cette thèse.

En conséquence de cette vision cognitive de la cohérence, il est possible d'adopter deux points de vue de celle-ci. Tantôt, on s'intéresse plutôt aux indices textuels qui sont interprétés par le lecteur en fonction de ses connaissances du monde et de la langue ; tantôt, on se concentre surtout sur les processus mis en jeu par le lecteur afin de construire une structure cohérente du texte, par exemple à la manière décrite par Kintsch et van Dijk (1978). Dans ce dernier cas, le processus principal mis en œuvre est l'inférence (Mason et Just, 2004), dont nous ne parlerons que brièvement ici, puisque la première optique correspond évidemment mieux aux objectifs de la lisibilité.

La cohérence et la compréhension

Après avoir précisé les notions de cohésion et de cohérence, reste à déterminer si la cohérence d'un texte influence bien la compréhension qu'un lecteur peut en avoir. Les études qui

traitent de cette question se focalisant tantôt sur la cohérence locale, tantôt sur la cohérence globale, nous avons reproduit ici cette distinction.

La cohérence globale est associée par Armbruster (1984, 203) aux « caractéristiques textuelles qui facilitent l'intégration des idées principales au sein d'une section, d'un chapitre, etc. », c'est-à-dire le système d'arrangement des idées ou le plan. À ce propos, Meyer *et al.* (1980) effectuent des expériences qui révèlent que reconnaître le plan utilisé par l'auteur d'un texte aide des étudiants de secondaire à mieux retenir son contenu, aussi bien au niveau des idées principales que des détails. Les auteurs observent aussi que les sujets capables de réutiliser le plan du document pour structurer leur compréhension sont ceux qui obtiennent les meilleurs résultats à des tests de compréhension standardisés.

À la suite d'une étude similaire effectuée sur des élèves de primaires (4e et 6e), Taylor (1980) constate que, après un délai de deux jours, le rappel des 6e années est supérieur lorsque ces derniers réemploient la structure du texte lors de la tâche de contrôle. Enfin, Slater *et al.* (1988) réalisent une expérience dans laquelle ils comparent la rétention d'étudiants auxquels on a enseigné des techniques de résumé, y compris l'explicitation de la structure, avec ceux de groupes contrôles ayant reçus un enseignement plus traditionnel. Ils concluent à un effet significatif de cet enseignement explicite des techniques de structuration textuelle dans la plupart des cas.

En ce qui concerne la situation en L2, Carrell (1984) effectue des expériences sur le taux de rétention en fonction du type de plan. Pour ce faire, elle modifie un texte afin d'en obtenir quatre versions différentes, chacune organisée selon l'une des quatre structures suivantes : collection de descriptions, cause/conséquence, problème/solution et comparaison. Ces documents sont ensuite soumis à des sujets dont le rappel immédiat est testé à l'aide d'un protocole écrit. L'auteur en conclut les remarques suivantes : (1) comme pour les natifs, certains types de plan semblent plus améliorer le rappel que d'autres ; (2) on observe des inférences avec les structures rhétoriques de la langue maternelle, bien que la mesure dans laquelle il est possible de transférer un schéma d'une langue à l'autre reste obscure ; (3) les apprenants qui emploient le bon plan sont plus performants, comme c'est le cas chez les natifs.

Au vu de ces quatre études, il apparaît que (1) si un texte est bien structuré et (2) si son lecteur est capable de reconnaître et d'utiliser cette structure, celui-ci intègrera alors mieux le contenu du texte qu'un lecteur qui n'a pas su utiliser cette structure. Toutefois, elles ne disent rien des effets des textes mal structurés sur les lecteurs. Une étude de Kintsch *et al.* (1977) nous éclaire à ce sujet. Elle compare les performances de sujets devant résumer deux extraits du *Décaméron* de Boccace dont les paragraphes sont tantôt présentés dans leur ordre normal, tantôt en désordre. On remarque que (1) les sujets confrontés aux fragments

mélangés prennent plus de temps pour les lire, ce qui laisse supposer que c'est à ce moment-là qu'est construite la macrostructure du texte, mais (2) ils mettent moins de temps pour les résumer. Pour les auteurs, ce serait une conséquence du temps supplémentaire passé à comprendre le texte lors de la lecture. De plus, (3) les experts chargés de discriminer entre les résumés réalisés à partir des extraits ordonnés et mélangés n'ont pas été capables de distinguer ceux relatifs à l'extrait 1, lequel est mieux structuré que le second. Pour ce dernier, par contre, leurs performances sont meilleures, en particulier lorsque le temps de lecture accordé aux sujets était limité. Cela montre que le degré de structuration du texte influence la lecture, puisqu'il a été plus aisé de reconstituer la macrostructure des textes désordonnés lorsque la cohérence de la version originale était plus grande.

On observe des résultats assez similaires en ce qui en ce qui concerne la cohérence locale, bien que le tableau soit plus complexe. Une première série d'études avancent que la cohésion référentielle diminuerait le temps de lecture et améliorerait la rétention (de Villiers, 1974; Kintsch *et al.*, 1975). Pour obtenir ces deux effets, de Villiers (1974) a utilisé des phrases simplement connectées via un article défini plutôt qu'un article indéfini. Par la suite, Bradshaw et Anderson (1982, cités par Myers *et al.*, 1987, 453) confirment que le rappel d'une phrase s insérée dans le contexte de deux autres phrases g et d est meilleur si s entretient un lien causal avec g et d que si elle n'en entretient aucun. La présence de liens cohésifs entre phrases produirait donc une lecture plus rapide et une meilleure compréhension.

Certains chercheurs (Keenan *et al.*, 1984; Myers *et al.*, 1987) ont toutefois souligné l'aspect binaire de ces expérimentations, qui n'opposent à des phrases cohérentes que d'autres, incohérentes. Keenan *et al.* (1984) s'interrogent plutôt sur les effets de l'intensité de la relation causale. Ils conçoivent des séries de quatre paires de phrases dans laquelle la conséquence reste la même et est précédée de quatre phrases différentes, avec lesquelles elle entretient une relation causale de plus en plus forte. Leurs résultats sont assez surprenants : si le temps de lecture diminue à mesure que la force cohésive augmente, le rappel suit une fonction quadratique en forme de U inversé. Autrement dit, ce sont les phrases modérément connectées qui sont les mieux remémorées. Myers *et al.* (1987) reproduisent la même expérience avec davantage d'items et en contrôlant plus strictement certains facteurs tels que la longueur des phrases et le nombre d'idées et ils obtiennent les mêmes résultats que Keenan *et al.* (1984).

L'existence de cette fonction quadratique intrigue et ces auteurs l'expliquent de la manière suivante. Pour deux phrases modérément connectées, il est généralement nécessaire que le lecteur effectue une inférence afin de rendre explicite leur lien à peine esquissé dans le texte. Ce processus de génération d'inférence explique un temps de lecture supérieur, mais produit un réseau de connexions plus riche entre les représentations des deux phrases en mémoire, facilitant leur rappel. Par contre, lorsque le lien entre les deux propositions est évident, la

production de cette inférence n'est pas nécessaire. Le temps de lecture diminue alors, tout comme la richesse des connexions entre les deux phrases en mémoire, d'où une moindre rétention.

Cette hypothèse a été éprouvée par Mason et Just (2004) via une analyse par IRMf (*cf.* Chapitre 1.2.3). Les auteurs ont fait lire les phrases de Myers *et al.* (1987) à des sujets, tandis que les zones activées de leur cerveau étaient cartographiées. Parmi celles-ci, les auteurs distinguent les aires du langage de l'hémisphère gauche (LH), celles de l'hémisphère droit (LD) et le cortex préfrontal dorsolatéral (CPFDL). Pour les différents niveaux de cohésion, on retrouve une activation de CPFDL qui croît à mesure que la cohérence entre paires de phrases décroît. Ce motif d'activation correspond à l'allongement du temps de lecture des phrases moins cohérentes observé par Myers *et al.* (1987) et serait lié à la phase de génération des inférences, qui mobilise d'autant plus de ressources que le texte manque d'indices de cohésion. Parallèlement, l'activation de LD suit la même forme en U renversé que le rappel chez Myers *et al.* (1987). Cette partie droite du cerveau serait responsable de l'intégration des inférences générées par le CPFDL au sein de la représentation interne du texte que s'est construit le lecteur via des processus contenus dans LH.

En résumé, ces résultats confirment l'idée que plus un texte est cohérent, plus il est compris facilement. Par contre, les effets de la présence de marqueurs de cohésion, qui sont justement des éléments que l'on peut mesurer le plus aisément en lisibilité, sont plus mitigés. Ainsi, lorsque la structure est trop aisée à reconstituer, le taux de rétention diminue. Alderson (2000, 68), va même jusqu'à penser que les effets de la cohésion sur la compréhension et sur la mémoire seraient faibles, car les lecteurs sont en général capables de faire des inférences, même en cas d'absence de marqueurs cohésifs. Plus optimistes, Sanders et Noordman (2000) estiment que la présence de marqueurs joue un rôle lors de la lecture, mais qu'elle n'affecte pas la structure textuelle construites par les lecteurs. Par contre, Degand et Sanders (2002) parviennent ensuite à une conclusion inverse en recourrant à des questions de compréhensions plutôt qu'au rappel. Il apparaît que les lecteurs de textes présentant des marqueurs ont mieux compris ceux-ci. Par ailleurs, ces auteurs rapportent un autre fait intéressant : les lecteurs en L2 ne semblent pas profiter plus ou moins des indices de cohérence que les lecteurs en L1.

On voit donc qu'il existe aujourd'hui encore plusieurs positions théoriques contradictoires concernant l'effet des marqueurs sur la compréhension. Ces divergeances viendraient notamment, pour Degand et Sanders (2002), des différentes techniques utilisées pour mesurer la compréhension. Quoi qu'il en soit au niveau théorique, de plus en plus d'études en lisibilité introduisent des facteurs de cohésion, cherchant ainsi à combiner les approches classiques et structuro-cognitives. Le point suivant passe en revue les différentes mesures

utilisées dans ce but.

Mesures de la cohérence

En ce qui concerne l'emploi de variables de cohésion ou de cohérence pour prédire la lisibilité de textes, on peut faire remonter les premières tentatives à Bormuth (1969a). Celui-ci estime qu'il est nécessaire de résoudre correctement les relations anaphoriques pour comprendre un texte et définit douze variables basées sur cette notion. Se basant sur une analyse de Menzel (1969, cité par Bormuth, 1969a, 38), il détermine 10 classes d'anaphores, dont le repérage est opéré manuellement au sein de son corpus. Il en dérive 10 variables, chacune correspondant au nombre d'anaphores d'un type en particulier sur le nombre total d'anaphores dans le texte. Par ailleurs, il compte aussi la densité des anaphores (exprimée en nombre par mot) et la distance moyenne (en nombre de mots) entre chaque anaphore et son antécédent. Ce sont ces deux dernières variables qui se révèlent les plus corrélées à la difficulté (respectivement, $r = 0,532$ et $r = -0,392$). Pourtant, les deux variables finalement sélectionnées pour les formules complexes font partie des 10 types. Il s'agit des répétitions référentielles (ex. **Johan** aime les pommes. **Johan** a mangé le gâteau aux pommes.) et des anaphores de classes (Le **cheval** s'ébroue. **Cet (animal)** est ...).

Par la suite, plusieurs autres techniques de paramétrisation des phénomènes de cohérence sont employées. Il est possible de classer la majorité d'entre elles en deux tendances : celles qui reposent sur le modèle propositionnel de Kintsch et van Dijk (1978) et celles qui recourent à la LSA. Par ailleurs, certaines approches échappent à cette classification, en particulier dans les travaux les plus récents. Nous détaillons ci-dessous ces trois approches.

Le modèle propositionnel et les inférences : Dans le chapitre 2.2.6, nous avons présenté le modèle propositionnel de Kintsch et van Dijk (1978). Sur cette base, Kintsch et Vipond (1979, 344-345) proposent une manière simple de paramétriser la cohérence locale d'un texte : ils considèrent que deux propositions sont connectées si elles partagent un même argument [17]. Haviland and Clarke (1974, cités par Kintsch et Vipond, 1979, 345) ont ainsi montré que des sujets comprennent plus vite une phrase lorsque celle-ci est connectée à la précédente via une répétition d'argument que lorsque ce n'est pas le cas. Bien sûr, lorsqu'on suit cette définition, certaines paires de phrases ne sont pas connectées. Pour les relier, il est alors nécessaire que le lecteur effectue une inférence, dont la production, comme il l'a été dit précédemment, ralentit la lecture, mais peut améliorer la rétention.

Cette dernière dimension (nombre d'inférences) est employée comme variable par les deux

17. Les auteurs reconnaissent que cette approche « n'est pas linguistiquement satisfaisante, mais elle suffit à leur objectif. Elle est simple, complètement objective et semble expliquer une partie importante de la variance dans la cohésion des textes » (Kintsch et Vipond, 1979, 345)

auteurs, sans qu'ils en testent expérimentalement l'efficacité. Dans des expériences postérieures, Kintsch (1979, 9-10) met au jour la faible corrélation que ce facteur entretient avec la difficulté des textes de son corpus. C'est surtout le nombre de réinsertions [18] par 100 mots qui est associé à la complexité textuelle ($r = -0,62$). Toutefois, ces deux variables sont reprises dans sa « formule de lisibilité ».

Dans la lignée de ces travaux, Clark (1981) suggère une approche simplifiée de l'analyse propositionnelle qui lui permet de prendre en compte à la fois nombre d'inférences requises par un texte, mais aussi la distance entre les pronoms et leurs référents, un aspect qui n'est pas sans rappeler la vision de Bormuth (1969a). Cependant, il ne rapporte pas de résultats concrets concernant l'efficacité de ces deux variables.

L'analyse sémantique latente : Ce second paradigme se développe en réponse au modèle de Kintsch et van Dijk (1978), que Foltz *et al.* (1998) estiment trop fastidieux, d'autant qu'il demande une longue formation afin d'être appliqué correctement. Ces auteurs recommandent à la place l'analyse sémantique latente (cf. Chapitre 2.2.8) comme un moyen plus aisé et automatisé de paramétriser la cohésion d'un texte. Celle-ci repose sur la similarité sémantique d'unités textuelles telles que les phrases ou les paragraphes. Elle capture dès lors la répétition lexicale, mais aussi des relations de synonymie et d'hyponymie. Par contre, elle ignore la plupart des autres types de noeuds cohésifs décrits par Halliday et Hasan (1976) (ellipse, référence pronominale, substitution, conjonctions causales).

La variable que Folz *et al.* utilisent est la moyenne des cosinus entre les vecteurs de chaque paire de phrases adjacentes. Ils en comparent les résultats avec des mesures de compréhension obtenues par Britton et Gülgöz (1991) et basées sur le nombre de propositions remémorées. Les corrélations entre ces deux techniques sont très élevées (r entre $0,98$ et 1), ce qui valide l'emploi de cette technique comme mesure de la cohésion [19]. Par contre, les auteurs n'évaluent pas son apport à une formule de lisibilité.

La LSA est réutilisée comme mesure de la cohérence textuelle au sein du logiciel Coh-Metrix (Graesser *et al.*, 2004). Les concepteurs offrent à l'utilisateur différentes variantes de cette technique. Au-delà de l'approche classique, proposée par Foltz *et al.* (1998) et qui est ici appelée LSA locale, on y trouve aussi la LSA globale (c'est-à-dire le cosinus moyen entre toutes les paires de phrases du texte) ; des calculs de similarité de chaque phrase avec son paragraphe et avec l'ensemble du texte, ainsi que des indices de la similarité de chaque

18. Comme nous l'avons déjà expliqué, les réinsertions ou réinstallations consistent à explorer la mémoire à long terme pour y chercher des structures précédemment rencontrées auxquelles raccrocher la structure en train d'être lue.

19. Notons toutefois que l'effet de la cohérence sur la compréhension est surtout perceptible dans le cas de lecteurs peu au fait du sujet traité dans le texte abordé. Ceux qui le maîtrisent déjà tirent plutôt parti de textes faiblement cohérents. Bien qu'utile, cette dernière donnée est toutefois difficilement modélisable sans un modèle de l'utilisateur.

paragraphe avec chacun des autres et l'ensemble du texte. Une fois de plus, aucune analyse de l'efficacité de différentes métriques en lisibilité n'est rapportée. Il faut attendre Pitler et Nenkova (2008) pour que la corrélation entre le cosinus moyen entre phrases adjacentes et la difficulté de textes soit évaluée. Le résultat est cependant peu encourageant, puisque la corrélation obtenue s'avère non significative ($r = -0,1012, p = 0,5947$)[20].

Parallèlement à la LSA, une mesure assez similaire a également été explorée par ces mêmes auteurs : le « word overlap », que l'on pourrait traduire par recouvrement lexical. Comme dans la LSA, les vecteurs de deux phrases sont comparés, mais il s'agit cette fois des vecteurs de mots originels et non ceux obtenus après une décomposition en valeurs singulières (SVD). Cette variante fournit dès lors une mesure plus directe du recouvrement lexical. Foltz *et al.* (1998) avaient comparé cette dernière avec la LSA et n'avait guère noté de différence (r compris entre $0,96$ et 1). Ils expliquent ces résultats similaires par une plus forte sensibilité de la LSA à la répétition lexicale entre unités qu'à la similarité plus indirecte qui est aussi calculée.

S'inspirant de cette idée, Graesser *et al.* (2004, 199) en implémentent trois variations dans Coh-Metrix : le « noun overlap » (nombre de noms en communs entre deux phrases) ; l'« argument overlap » (quand une phrase s_i comporte un nom, dont la racine est commune avec celle d'un nom d'une phrase s_j) et le « stem overlap » (quand un nom de s_i a une racine commune avec un mot d'une quelconque catégorie de discours de s_j). Chacune de ces variantes est calculée comme suit : on définit une matrice $\mathbf{R}$ de dimensions $n * n$ (n est le nombre de phrases dans le texte) où chaque cellule comprend la valeur 1 ou 0 selon qu'il y ait répétition lexicale entre les deux phrases ou non. À partir de cette matrice, il est possible de calculer diverses mesures de la répétition lexicale entre phrases selon les cellules prises en compte. Ainsi, le « noun overlap » est défini comme suit :

$$\text{Co-reference cohesion local-1} = \frac{\sum_{i=1}^{n-1} R_{i,i+1}}{n-1} \tag{3.9}$$

Une mesure de cohérence globale est également définie :

$$\text{Co-reference cohesion global} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n-1} R_{i,j} \mid i < j}{n * \frac{n-1}{2}} \tag{3.10}$$

Entre ces deux extrêmes, il est possible d'associer à $\mathbf{R}$ une matrice de poids, qui pondère l'importance de la répétition lexicale en fonction de la distance entre entre chaque paire de phrases. Cependant, aucune évaluation de ces différentes variables n'est effectuée, si ce n'est par Crossley *et al.* (2008), qui prennent en compte le « content word overlap », c'est-

20. Le corpus d'évaluation compte ici trente textes et non plus quatre comme chez Foltz *et al.* (1998).

à-dire le recouvrement des mots de contenu entre phrases adjacentes. Cette variable obtient un taux de corrélation ($r = 0,79$) sur les passages de Bormuth, dont on se souviendra qu'ils produisent toujours des coefficients de corrélations élevés.

Plus tard, Pitler et Nenkova (2008) obtiennent par contre des corrélations non signficatives aussi bien pour le recouvrement lexical ($r = -0,0531$, $p = 0,7806$) que pour le recouvrement des noms et des pronoms uniquement ($r = 0,0905$, $p = 0,6345$). En conséquence, l'efficacité de la LSA et des différentes mesures de recouvrement lexical entre phrases restent sujettes à caution. De plus, comme ces mesures n'ont jamais été explorée en lisibilité des L2 et qu'elles reposent sur l'un des rares modèles sémantiques qui peut être développé sans ressources linguistiques riches, nous avons choisi de les explorer.

Autres mesures : En plus de ces deux tendances ancrées dans les travaux de Kintsch et de ses collaborateurs, il faut souligner quelques autres approches avant de clore cette section. À nouveau, le logiciel Coh-Metrix se révèle une riche source d'inspiration à ce niveau. Graesser *et al.* (2004) suggèrent un indice de la cohésion causale qui évalue dans quelle mesure les phrases d'un texte sont connectées au moyen de relations causales [21]. Celui-ci correspond au rapport entre P, le nombre de particules causales (ex. : parce que, en conséquence, etc.), et $V + 1$, où V est le nombre de verbes à valeur causale dans le texte. Cette dernière information est obtenue à l'aide de WordNet (Fellbaum, 1998).

Bien que ce prédicteur n'ait pas, à notre connaissance, été testé en lisibilité, McNamara *et al.* (2010) ont réalisé une étude analysant la capacité de diverses variables décrites dans Coh-Metrix à distinguer des textes fortement cohérents d'autres faiblement cohérents. Leurs expériences révèlent que la cohésion causale et la co-référence nominale sont les deux meilleurs facteurs discriminants, tandis que la LSA et le « stem overlap » sont moins utiles.

Un second type de mesures est celles basées sur les connecteurs présents dans les textes. (Graesser *et al.*, 2004) classent ceux-ci selon les 6 familles suivantes : négatif (ex. *but, although*) ; positif (ex. *and*), causal, additif (ex. *moreover*, *also*) ; temporel et logique (ex. *or, if*). Cependant, ils ne précisent pas comment ils détectent ces termes, ni comment ils s'en servent pour définir des variables. Il est probable qu'il s'agisse simplement d'un ratio.

Par la suite, Pitler et Nenkova (2008) proposent deux nouvelles familles de variables : la cohérence des entités et les relations de discours. Pour la première, ils s'appuient sur le modèle de Barzilay et Lapata (2005, 2008) qui proposent une alternative à la LSA pour évaluer automatiquement la cohésion de textes. Un texte y est représenté sous la forme d'une matrice reprenant les « entités de discours » présentes dans chaque phrase. Par « entité

21. Notons que cette mesure n'est valable que dans les textes où apparaissent ce genre de relations, tels que les textes scientifiques.

de discours », les auteurs désignent les syntagmes nominaux participant à une relation de coréférence et auxquels est associés une fonction (sujet, objet, autre ou absent de la phrase). Ensuite, un algorithme évalue le niveau de cohésion du texte sur la base des transitions entre entités.

C'est ce principe qui est repris par Pitler et Nenkova (2008, 191), lesquelles utilisent les probabilités de passer d'un type d'entité à un autre dans un texte comme autant de variables de lisibilité (17 au total). Cependant, aucune d'elles n'apparaît corrélée significativement avec la difficulté. Cette procédure a été répliquée par Feng *et al.* (2009, 2010) au sein d'une formule par régression logistique, laquelle atteint une précision de classification nettement inférieure à une formule basée sur le même modèle, mais qui utilise simplement le nombre moyen de mots par phrase et le nombre moyen de syllabes par mots ($42,14\%$ contre $53,07\%$). Au-delà de ces résultats insatisfaisants, la technique requiert un parseur syntaxique et d'un résolveur d'anaphores afin d'obtenir la matrice des entités, ce qui rend sa mise en place extrêmement complexe.

Le second ensemble de facteurs exploré par Pitler et Nenkova (2008) est basé sur les relations de discours et s'inspire des modèles de langue. Chaque texte est modélisé comme un sac de relations (et non un sac de mots). Celles-ci expriment par exemple une comparaison, une relation temporelle, etc., et peuvent être explicites, si elles sont marquées par un connecteur, ou implicites dans le cas contraire. Une fois constitué ce sac de relations, on en calcule la probabilité, non pas à l'aide d'un modèle n-grammes, mais d'une distribution multinomiale sous forme logarithmique (Pitler et Nenkova, 2008, 192) :

$$\log(P(n)) + \log(n!) + \sum_{i=1}^{k}(x_i \log(p_i) - \log(x_i!)) \tag{3.11}$$

Les auteurs avancent trois arguments en faveur de cette distribution par rapport à un modèle unigramme : elle prend en compte la longueur des textes, elle évite le problème de la division par 0 lorsqu'un texte ne comprend pas de relations et elle capture les motifs relatifs de fréquence. Ils obtiennent d'ailleurs des corrélations significatives pour cette variable ($r = 0,4835$, $p = 0,0068$), qui est aussi leur meilleur prédicteur, ainsi que pour le nombre de relations par texte normalisé selon le nombre de mots ($r = 0,3819$, $p = 0,0373$)[22]. Malheureusement, la procédure en sac de relations n'étant actuellement pas automatisée, cette variable ne peut être insérée au sein d'une formule automatique.

Terminons les mesures de cohérence par les travaux de Feng *et al.* (2009, 2010) qui ont

22. Par contre, ni le nombre de relations non normalisé ($r = -0,2729$, $p = 0,1445$), ni les probabilités des relations implicites ($r = 0,2403$, $p = 0,2009$) ou explicites ($r = 0,1528$, $p = 0,4203$) seules ne se révèlent intéressantes.

appliqué les chaînes lexicales décrites par Galley and McKeown (2003, cités par Feng *et al.*, 2010, 278) à la problématique de la lisibilité, dans le but de modéliser le nombre total de concepts qu'un lecteur doit avoir à l'esprit à un moment donné. Une chaîne lexicale correspond à un ensemble de noms, présents dans un même document, qui entretiennent des relations de synonymies, d'hyponymies, etc. Une fois ces chaînes reconnues au terme d'un traitement automatique, les auteurs en tirent les six variables suivantes : la longueur des chaînes d'un texte mesurée tantôt en nombre de concepts (1), tantôt en nombre de mots entre la première entité et la dernière (2) ; le nombre moyen de chaînes actives à chaque mot (3) ou à chaque syntagme nominal (4), le nombre de chaînes dans le document (5) ou encore le nombre de chaînes dont la longueur dépasse la moitié du texte (6). Cependant, une fois ces différentes variables combinées dans un modèle logistique (*cf.* Chapitre 5.3.3), elles ne se révèlent pas plus prédictives que celles basées sur la cohérence des entités ($42,58\%$).

Conclusion

On voit que de nombreuses tentatives ont été effectuées pour prendre en compte les divers aspects de la cohésion et de la cohérence en lisibilité et qu'il en ressort deux aspects : la prépondérance de la microstructure et la faible capacité de prédiction. La première de ces deux caractéristiques n'est pas étonnante, puisqu'il est plus aisé de modéliser automatiquement des aspects de la cohésion interphrastique que de détecter par voies informatisées la structure textuelle d'un texte entier. En ce qui concerne la capacité de prédiction de ce type de variable, on aurait pû s'attendre à des merveilles, au vu de l'enthousiasme des structuro-cognitivistes. Pourtant, la majorité des expériences rapportés précédemment, y compris les plus récentes, rapportent des effets peu ou prou significatifs, et souvent moins importants que ceux des prédicteurs classiques. On notera les exceptions constituées par les variables basées sur les relations de discours (Pitler et Nenkova, 2008) (qui requièrent malheureusement des modèles complexes, implémentés dans des logiciels qui font encore défaut en français, à notre connaissance) et le recouvrement des mots de contenu chez Crossley *et al.* (2008). Pour ces deux raisons, nous avons limité l'exploration des variables de cohérence au « word overlap » et à une approche par LSA.

3.2.6 Variables spécifiques à la lecture en L2

Jusqu'à présent, nous avons traité de facteurs dont la validité avait surtout été évaluée dans un contexte L1 et dont nous avons discuté la pertinence dans un contexte L2 à partir des données expérimentales disponibles. Or, si certains auteurs tels que Greenfield (2004) doutent de la nécessité de prendre en compte des aspects propres aux processus de lecture en L2, nous nous positionnons plutôt dans la mouvance de Tharp (1939), Laroche (1979) et Uitden-

bogerd (2005) qui défendent au contraire l'intérêt de modéliser ces dimensions spécifiques. Nous avons nous-mêmes montré, dans une étude préliminaire (François, 2009a), combien simplement envisager les temps et modes verbaux peut conduire à une amélioration significative des performances pour une formule de lisibilité en L2.

Par contre, comme le chapitre 2 l'a rappelé, cette optique n'a guère été explorée par les chercheurs en lisibilité des L2, qui se sont essentiellement concentrés sur un aspect particulier des effets d'interférence entre la L2 et la L1 : les congénères (Tharp, 1939; Uitdenbogerd, 2005). Dans cette section, nous résumons ces travaux sur les congénères, mais présentons également deux autres pistes qui nous ont semblées dignes d'intérêt : les étapes d'acquisition (aspects verbaux et vocabulaire) et les unités polylexicales. Si la première est simplement esquissée, notre étude explore la seconde nettement plus en profondeur que ce qui a été fait jusqu'alors. Nous terminons cette partie par un aspect qu'il a été difficile de ranger dans l'une de nos catégories : le type de textes.

Interférences avec la langue maternelle

Il est assez évident que les caractéristiques de la langue maternelle d'un lecteur vont influencer sa pratique en L2. Nous avons déjà souligné que la reconnaissance visuelle d'un mot est influencée par la proximité orthographique qu'il entretient avec les mots de l'autre langue (Font et Lavaur, 2004) ou encore par d'éventuels sens différents qu'il peut endosser dans la L1 (Degani et Tokowicz, 2010) (il s'agit dans ce cas de faux amis). Toutefois, il existe un autre phénomène d'interférences entre langues qui a fait l'objet de plusieurs études en lisibilité : les congénères ou cognats [23]. Il s'agit de mots d'une L2 dont la forme est homographique ou presque avec leurs équivalents sémantiques dans la L1 du lecteur.

Cet effet de facilitation des congénères a été confirmé par de nombreuses études sur la reconnaissance des mots bilingues (Caramazza et Brones, 1979; Dijkstra *et al.*, 1999, 2010; Lemhöfer *et al.*, 2008), mais aussi par une étude neuropsychologique reposant sur une tâche de prononciation à base d'images où l'activité du cerveau est observée via une TEP (De Bleser *et al.*, 2003). Ces auteurs ont ainsi montré que les aires cérébrales activées pour les mots en L1 et leurs congénères L2 sont les mêmes, mais que des aires additionnelles entrent en activité pour les mots de L2 qui ne sont pas des congénères. On peut ainsi conclure que l'effet de facilité des congénères vient d'un plus grand partage des structures neuronales incriminées.

D'ailleurs, l'effet de facilitation des congénères a souvent servi d'argument pour défendre la

23. Il s'agit apparemment d'un anglicisme, dérivé de son équivalent anglais *cognate*, lequel co-existe avec le terme « congénère ».

vision d'un lexique bilingue intégré et non pas d'un accès sélectif en fonction de la langue. Les parfaits congénères sont aujourd'hui vus comme partageant effectivement la majorité de leurs représentations entre les différentes langues connues par l'apprenant. Cependant, Font (2001, cité par Dijkstra *et al.*, 2010, 287), entre autres, a montré qu'un effet de facilitation était également observable, bien que moindre, pour les congénères voisins, c'est-à-dire qui diffèrent par une seule lettre. Plus généralement, Dijkstra *et al.* (2010) envisagent la proximité des congénères selon un continuum et rapportent pas moins de quatre tentatives différentes d'explication de cet effet de facilitation.

Enfin, Tharp (1939) soulignait déjà que cet effet facilitateur ne joue à la lecture que si l'équivalent en L1 du mot L2 à reconnaître est connu du lecteur. Cette hypothèse s'est vue confirmée et étendue par les résultats de Font (2001, cité par Dijkstra *et al.*, 2010, 287), qui révèlent que la facilité apportée par les congénères s'amoindrit à mesure que la fréquence du terme diminue. En résumé, on peut dire que l'effet de facilitation des congénères est fonction à la fois de la proximité des formes orthographiques et de la fréquence des mots.

Les chercheurs en lisibilité n'ont cependant pas attendu de connaître ces études pour insister sur l'importance des congénères (*cf.* Chapitre 2.3.2). Depuis, Arnold (1932), cette problématique est l'une des plus envisagées en lisibilité L2 (Spaulding, 1951; Tharp, 1939; Uitdenbogerd, 2005) et on peut dire qu'elle a fait l'objet de deux approches différentes.

La première d'entre elles consiste à adapter certaines des variables lexicales présentées précédemment, afin de prendre en compte l'effet facilitateur des congénères du texte à analyser. Nous avons déjà évoqué en ce sens l'indice de fréquence de Tharp (1939), repris par Spaulding (1951), où chaque congénère se voit attribuer la classe de fréquence la plus courante. La validité de cette stratégie semble avalisée par les résultats de Spaulding (1951), qui compare son indice ($r = 0,688$) à une variable de fréquence lexicale classique ($r = 0,62$). En réalité, il faudrait probablement tenir compte des congénères pour l'ensemble des variables lexicales, comme le défend Laroche (1979) en ce qui concerne la longueur des mots : « for a foreign student, a long cognate is more welcome than a short unrelated word ».

La seconde technique de paramétrisation est utilisée par Uitdenbogerd (2005, 21). Il s'agit simplement de considérer la proportion de congénères dans le texte à analyser. Ce faisant, l'auteur obtient, pour cette variable, un coefficient de corrélation élevé avec son critère ($-0,61$), lorsqu'elle ne considère que les textes ordonnés par des non-natifs. Par contre, cette association diminue nettement pour l'ordonnancement des natifs ($r = -0,33$).

Que l'on retienne la première ou la seconde approche, toutes deux sont difficilement applicables en raison des deux obstacles suivants. D'une part, il est nécessaire d'entraîner un modèle spécifique à chaque paire de langues (à part dans le cadre de deux langues fortement

éloignées, auquel cas le concept même de congénère devient improductif). Ceci demande non seulement de développer un extracteur automatique des congénères (cf. Frunza et Inkpen, 2007; Kondrak, 2004; Mitkov *et al.*, 2007), mais également de disposer d'un corpus adapté, c'est-à-dire pour lequel l'effet facilitateur qui résulte des congénères a été pris en compte. D'autre part, il faut aussi définir si l'équivalent L1 d'un congénère donné est connu par un lecteur. Ceci peut être toutefois être grossièrement estimé à l'aide des fréquences objectives des termes, comme nous l'avons indiqué précédemment. Nous n'avons donc pas pu expérimenter cette variable, étant donné notre volonté de développer un modèle indépendant d'une langue maternelle en particulier.

Les unités polylexicales

Il existe une autre facette textuelle tout aussi susceptible de représenter un bon prédicteur de la difficulté lexicale pour des apprenants de L2 : les collocations et expressions figées présentes dans le texte à analyser. Une bonne connaissance de celles-ci est en effet associée à une pratique fluide et appropriée de la langue (Pawley et Syder, 1983). On peut donc s'attendre à ce que les apprenants d'une L2, et en particulier les débutants, rencontrent des difficultés dans le traitement de ces chaînes lexicales. Pourtant, cette hypothèse n'a fait l'objet d'aucune étude poussée, ni pour le FLE, ni pour l'anglais L2.

Commençons par préciser ces différentes notions, vu que de nombreuses définitions du terme collocation ont été proposées dans la littérature, sans qu'aucune ne fasse complètement l'unanimité. Dans cette thèse, nous reprenons la distinction généralement acceptée entre combinaisons libres, collocations et expressions figées (Bahns et Eldaw, 1993). Dès lors, les collocations sont vues comme un élément lexical préférant la compagnie d'un autre terme plutôt que ses synonymes à cause de contraintes qui ne relèvent ni de la syntaxe, ni de la sémantique, mais de l'usage (Van Roey, 1990). Par ailleurs, leur sens global tend à refléter le sens de leurs éléments constitutifs, ce qui n'est généralement pas le cas des expressions figées (Bahns et Eldaw, 1993).

Une fois cette distinction posée, nous rassemblons les expressions figées et les collocations sous un hyperonyme commun : les UPs. En effet, nous envisageons ces deux unités à l'aide d'une définition statistique qui, dans la pratique, ne permet pas de les différencier. Dès lors, les UPs sont vues comme des « suites de mots qui se trouvent plus fréquemment associés qu'il ne le seraient par le seul fruit du hasard » (Dias *et al.*, 2000, 213), ce qui peut se mesurer à l'aide de mesures d'association (*cf.* Section 3.3.4).

Une autre raison en faveur de ce regroupement est que ces deux types d'entités lexicales semblent produire un effet similaire à la lecture. C'est du moins ce que rapportent diverses

expérimentations effectuées sur des natifs, mais aussi sur des non-natifs. Ces travaux s'appuient sur l'hypothèse que les UPs sont complètement ou partiellement stockées dans la mémoire à long terme (Pawley et Syder, 1983) et peuvent être récupérées telles quelles, ce qui soulage la mémoire à court terme dont les capacités limitées sont connues depuis longtemps (Miller, 1956). Dès lors, le traitement des UPs devrait être plus rapide à la lecture comme à la production orale, au moins par les natifs qui sont familiers avec la majorité d'entre elles.

Si cette hypothèse n'est pas encore considérée comme un fait avéré, plusieurs travaux (Conklin et Schmitt, 2008; Jiang et Nekrasova, 2007; Underwood *et al.*, 2004) ont effectivement observé qu'en moyenne, le traitement des UPs est plus rapide que celui des chaînes libres, à la lecture comme lors de la production orale. Ce qui est étonnant, c'est qu'un effet similaire a été observé pour des apprenants de niveau avancé. En effet, la maîtrise des UPs requiert d'avoir intégré un grand nombre de segments en mémoire et donc d'avoir déjà été confronté à une large quantité de données linguistiques, ce qui est rarement le cas des apprenants, surtout dans les niveaux inférieurs.

Il faut noter que ces études portent sur le temps de lecture et, donc, de reconnaissance des UPs, mais elles n'évaluent pas directement leur effet sur la compréhension. À ce sujet, Underwood *et al.* (2004) rapportent que certains apprenants, chez lesquels un traitement plus rapide des UPs a été observé, ignoraient le sens de près d'un tiers de celles-ci. Ils estiment que ces derniers ont sans doute déjà été en contact avec les chaînes lexicales considérées, ce qui expliquerait le déclenchement de l'effet facilitateur [24]. Toutefois, chez des apprenants débutants ou intermédiaires, il y a fort à parier que cet effet soit contrebalancé par le fait que les UPs rencontrées sont (1) majoritairement inconnues des lecteurs et (2) d'autant plus difficiles à élucider à l'aide du contexte que leur sens peut être non compositionnel.

Dans ce cas, un moyen pratique d'estimer si une unité polylexicale (UP) donnée est connue ou pas consiste à se baser sur leur fréquence d'apparition dans un large corpus. Cela revient à postuler qu'une UP plus rare est moins familière et partant, davantage susceptible d'être une source de problème à la lecture. Cette hypothèse a été formalisée par Weir et Anagnostou (2008) qui proposent de se servir de la moyenne des fréquences absolues des collocations présentes dans un texte (ACF) comme d'un indice de sa difficulté (Weir et Anagnostou, 2008) :

$$\text{ACF} = \frac{1}{n_c}(\sum_{i=1}^{m} f_i * n_i) \tag{3.12}$$

24. Rappelons également les résultats de Solomon et Postman (1952) qui ont observé que le gain de vitesse dans la reconnaissance est présent, même quand le sens des mots n'est pas compris.

Où n_c correspond au nombre total de collocations et m au nombre de collocations différentes dans le texte ; f_i représente leur fréquence absolue dans un corpus de référence et n_i leur fréquence absolue dans le texte à analyser. Le problème de cette étude, comme nous l'avons dit, est qu'elle ne rapporte aucun résultat validant cette approche.

Par contre, dans un article antérieur, Ozasa *et al.* (2007) avaient présenté une formule de lisibilité pour des apprenants japonais de l'anglais langue étrangère (ALE) qui comprend, entre autres variables, un indice de la difficulté des collocations. Celui-ci, qu'ils nomment « Difficulté des idiomes basés sur des manuels », a été calculé à partir d'une évaluation intuitive de professeurs japonais experts dans l'enseignement de l'ALE. Il apparaît que cette variable n'est pas significative au sein de leur modèle de régression linéaire multiple, puisqu'elle obtient un t [25] de $0,4987$, ce qui correspond à une p-valeur de $0,6188$ (Ozasa *et al.*, 2007, 4).

Ces deux études semblent relativement contradictoires. La première défend l'emploi des collocations au sein d'une formule de lisibilité L2, sans appuyer cette affirmation par des résultats tangibles, tandis que la seconde rapporte un effet négligeable de cette variable. Pour cette raison, nous avons décidé d'évaluer plus précisément l'apport des collocations en lisibilité du FLE. En effet, il apparaît que ce phénomène des collocations est loin d'être anecdotique dans la langue, puisqu'un tiers au moins de celle-ci serait constitué d'UPs (Conklin et Schmitt, 2008, 72).

Les étapes d'apprentissage

Une autre piste, que nous n'explorons pas dans cette étude, mais mériterait de l'être par la suite, sont les étapes d'acquisition. En effet, « la recherche sur l'acquisition des langues montre [...] que certains éléments d'une langue s'acquièrent dans un ordre spécifique » (Schlyter, 2003, 1). C'est vrai en ce qui concerne l'acquisition de la langue maternelle, mais apparemment aussi en ce qui concerne l'acquisition des L2, et cela, indépendamment de la technique d'apprentissage (scolaire et guidée ou plus naturelle et en contexte). Dès lors, il devrait être possible de se baser sur ces différents stades et leurs caractéristiques pour évaluer la lisibilité de textes.

Un premier problème à cette démarche est que les différents stades et leurs contenus sont décrits, du moins dans l'étude de Schlyter (2003), à partir de productions orales d'apprenants. Or, si comme le défend Pienemann (1998) l'enseignement d'un phénomène linguistique ne produit d'effets qu'une fois que l'apprenant est « mûr », comment définir ces mêmes stades

25. Dans le contexte de la régression linéaire multiple, la statistique t résulte d'un test de significativité sur le coefficient d'une des variables explicatives (Howell, 2008, 264).

pour la compréhension ? D'autant qu'il est plus difficile de mesurer une difficulté précise à la compréhension qu'à la production. On pourrait aussi se reposer sur des stades de production en considérant qu'un apprenant n'ayant pas encore la maîtrise d'une forme grammaticale en production éprouvera également plus de difficulté à la comprendre.

Une autre solution est apparue récemment sous la forme de référentiels en cours de publication chez Didier et rédigés par une équipe internationale chapeautée par Beacco et Porquier. Ces guides pour les professeurs s'accordent avec le CECR et listent, pour chaque niveau, une série de termes, de structures qui devraient être maîtrisées par un apprenant parvenu à ce stade. Malheureusement, ces guides n'ont paru que récemment et ne couvrent pas encore l'ensemble des six niveaux du CECR. C'est pourquoi, bien qu'il s'agisse d'une perspective très intéressante, nous n'avons pas pu l'explorer dans le cadre de ce travail.

Le type de texte

La dernière famille de variable que nous traiterons ici est le type de texte. Il peut paraître étonnant de classer cet aspect dans les facteurs propres au contexte L2 plutôt que sous la section sur la cohérence, mais il nous a semblé y avoir des raisons pour ce faire. En consultant des manuels de FLE, nous avons observé que la répartition des types de textes varie nettement en fonction de la difficulté (*cf.* Chapitre 4.3.5). En particulier, de nombreuses méthodes privilégient les dialogues dans les premières étapes de l'apprentissage, avant de passer à des textes narratifs ou informatifs.

Cependant, il semblerait que cette observation dépasse le cadre restreint des manuels de FLE, puisqu'elle avait déjà été avancée pour l'anglais L1 par plusieurs chercheurs et le français L1 par Henry (1975). Dès les années 40, Dolch (1948, 240) suggère que le type de discours peut constituer une variable intéressante. Il considère que les textes non-narratifs sont plus complexes que les textes narratifs, notamment parce que ces derniers comportent des conversations. Or, les conversations constitueraient des fragments textuels plus simples pour quatre raisons : les idées exprimées sont plus proches des idées quotidiennes et sont aussi plus concrètes ; les phrases sont plus simples et plus courtes et les termes employés sont plus familiers. Toutefois, il renonce à utiliser cette variable, disant que « The per cent of conversation seemed an indirect rather than a direct way of measuring reading difficulty. If more conversation made for easier reading, what elements in conversation made it easier ? ».

Par la suite, Flesch (1948, 227) et Gunning (1952, 32) estiment eux aussi qu'un dialogue est plus lisible et plus attractif qu'un autre type de texte. Quant à Henry (1975), il expérimente l'efficacité de variables supposées identifier un dialogue. Il en définit trois : le pourcentage de pronoms personnels de la 3^{e} personne ; le pourcentage de pronoms personnels de dia-

logue (1re et 2e personnes) [26] et le nombre d'indicateurs de dialogue (catégorie qui reprend les points d'exclamation, les guillemets et les prénoms employés seuls). On ne connaît pas exactement leur efficacité, car l'auteur ne rapporte pas de corrélations. Toutefois, l'introduction de l'un ou l'autre des deux derniers facteurs dans toutes les formules d'Henry laisse présager de leur pertinence. C'est pourquoi nous avons décidé d'implémenter des variantes de ces facteurs.

Au-delà de la simple distinction entre les dialogues et les autres types de textes, il serait possible d'envisager de détecter plus globalement les différents types de textes (narratif, descriptif, informatif, dialogue, etc.), afin de déterminer si leurs propriétés ne sont pas caractéristiques de différents niveaux de difficulté. Toutefois, les outils de segmentation automatique en type de texte, tel celui de Cocco *et al.* (2011), sont encore en voie de développement et nous n'avons pas pu envisager cette perspective dans ce travail, nous limitant à une approche plus simple basée sur le repérage des dialogues.

Conclusion

Au terme de ce long parcours théorique sur les indices de lisibilité, il est possible d'en retirer plusieurs observations d'intérêt qui nous semblent remettre en question certaines positions défendues précédemment dans la littérature. Tout d'abord, nous avons montré qu'il existe aujourd'hui un large spectre de facteurs utilisables en lisibilité et que ceux-ci relèvent de dimensions textuelles variées. Même si certaines de ces dimensions restent encore inexplorées (par exemple, les aspects pragmatiques et les connaissances *a priori* de l'apprenant, en particulier sa langue maternelle), on est loin de l'opposition facteurs classiques / facteurs structuro-cognitivistes qui régnait dans les années 80. Les études les plus récentes dans le domaine (Feng *et al.*, 2010; Pitler et Nenkova, 2008) ont montré que l'utilisation du TAL permet de réunir ces deux paradigmes au sein d'une méthodologie unique qui cherche à profiter des avantages de ces deux démarches.

En analysant les résultats de ces synthèses, mais aussi ceux d'autres études plus anciennes, l'un des aspects qui étonnent le plus est l'efficacité jamais démentie des prédicteurs classiques. Autant la critique des structuro-cognitivistes avait été rude à leur encontre, autant la plupart des chercheurs qui ont pris la peine de comparer ces facteurs et d'autres qu'ils estimaient plus pertinents se sont rendus compte qu'il n'était pas si évident de trouver mieux que les bonnes vieilles recettes. Ainsi, à combien de reprises avons-nous remarqué que de nouvelles variables – en particulier celles issues du paradigme structuro-cognitiviste et celles

26. Le lecteur attentif aura noté que ces mesures ont également été utilisées pour mesurer le taux de personnalisation des textes. Cela nous indique que certaines variables peuvent être interprétées diversement, comme c'est aussi le cas pour les ratios de catégories de discours.

traitant de la complexité syntaxique – conduisaient à des performances moindres qu'une simple formule prenant en compte le nombre de mots par phrase et le nombre de lettres par mot ? Ce fait est d'autant plus regrettable que la plupart de ces nouvelles variables reposent sur des modèles théoriques complexes, qui requièrent un traitement informatisé nettement plus lourd.

On peut avancer deux raisons différentes à ces faibles capacités prédictives : soit une déficience au niveau de l'hypothèse qui sous-tend la variable, soit un problème au niveau de la paramétrisation elle-même. Le premier cas apparaît lorsque la méthode de paramétrisation est fiable et produit une annotation des données correcte. Dans ce cas, un résultat médiocre nous indique effectivement que l'information considérée n'est pas corrélée avec la difficulté et ne constitue donc pas un bon prédicteur. Toutefois, l'intégration d'outils TAL en lisibilité peut conduire à une seconde source d'inefficacité : une paramétrisation incorrecte ! En effet, il ne suffit pas de produire une annotation reproductible et objective (ce que Harrisson (1980) désigne comme la fiabilité de l'analyste). Encore faut-il que celle-ci corresponde bien aux connaissances définies dans le modèle théorique, sans quoi l'hypothèse de départ peut très bien être correcte, mais fournir de mauvais résultats simplement parce que la procédure de paramétrisation commet trop d'erreurs.

En d'autres termes, la supériorité de certaines variables plus simples pourrait s'expliquer par des erreurs dans la paramétrisation de leurs concurrentes plus complexes. Ainsi, on ne s'étonnera pas que l'un des meilleurs prédicteurs de la cohérence chez Pitler et Nenkova (2008), à savoir la probabilité des relations de discours d'un texte, repose sur une annotation manuelle et que d'autres dimensions similaires, basées sur des procédures automatisées, produisent des résultats nettement inférieurs.

Un troisième aspect qui mérite d'être relevé est la possibilité de fonder théoriquement la pertinence de la plupart des variables que nous avons recensées. On a souvent reproché aux formules de lisibilité de se limiter à des indices de la difficulté, sans employer directement des facteurs causaux. Pourtant, pour la plupart de nos facteurs, nous avons trouvé des preuves d'un tel lien causal avec la difficulté du côté de la psychologie expérimentale. Nous avons montré que bien des variables classiques sont en fait tout à fait justifiées théoriquement, à la différence de ce qu'ont pensé nombre de structuro-cognitivistes. Parmi les facteurs importants, il n'y a guère que la longueur moyenne des phrases pour lequel un effet spécifique n'a pas pu être identifié. En effet, les études en psychologie de la lecture ont bien confirmé l'existence d'un effet de fréquence, de répétition, de l'influence des structures syntaxiques ou encore du taux de cohérence. Nous pensons donc que le pont que nous nous sommes efforcé de construire entre ces deux disciplines mérite l'attention. Il laisse présager une lisibilité qui s'attaquerait aux causes réelles de la difficulté à la lecture

Soulignons enfin les apports de cette étude en ce qui concerne les facteurs eux-mêmes. Nous en avons ainsi identifié plusieurs qui n'ont soit jamais été considérés dans le contexte du FLE (citons parmi les plus importants, les aspects verbaux, la densité conceptuelle et les aspects de cohérence), soit jamais envisagés tout court (nombre de voisins orthographiques, structures des unités polylexicales, le TTR normalisé, les étapes d'acquisition ou le nombre de sens d'un mot). Malheureusement, pour des raisons que nous avons précisées à chaque fois, il n'a pas été possible d'implémenter l'ensemble de ces variables dans le cadre de cette thèse. Celles que nous avons retenues sont décrites dans la section suivante.

3.3 La paramétrisation des variables retenues pour notre modèle

Dans cette seconde partie, nous présentons la liste des variables que nous avons retenues pour nos expérimentations. Celles-ci visent un double objectif : apporter un éclairage expérimental sur les discussions théoriques débattues dans ce chapitre et déboucher sur un modèle optimal pour une première formule de lisibilité computationnelle pour le FLE. Au vu de ce double contexte, il est évident qu'il s'agit de tester un maximum de facteurs et, bien que nous n'ayons pas eu la possibilité d'implémenter l'ensemble des prédicteurs identifiés dans la partie précédente, notre liste en comporte tout de même 406. Ceux-ci relèvent de quatre niveaux différents – les aspects lexicaux, syntaxiques, conceptuels et spécifiques au FLE – et sont classés en fonction de la nomenclature établie dans la première partie de ce chapitre.

Pour chaque famille de prédicteurs considérée, les différentes paramétrisations possibles sont discutées et les détails d'implémentation sont décrits. Dans la majorité des cas, nous avons refusé de nous limiter à un ensemble de paramètres considéré optimal sur la base d'arguments théoriques et avons implémenté les différentes variantes d'un même indice. C'est notamment ce qui explique le nombre élevé de variables. On notera qu'un identifiant unique est associé à chaque variable afin de faciliter la discussion des résultats qui a lieu dans le Chapitre 5.

3.3.1 Les variables lexicales

C'est certainement au niveau des prédicteurs lexicaux que notre étude est la plus complète. Parmi ceux-ci, l'effet de fréquence est l'un des plus étudiés, aussi bien en psychologie cognitive qu'en lisibilité. De nombreuses paramétrisations en ont été envisagées dans la littérature, sans qu'aucune ne fasse pleinement l'unanimité, d'où notre volonté d'en expérimenter autant que faire se peut. Nous avons ainsi implémenté pas moins de 78 statistiques de fréquence lexicale, 26 variantes du nombre d'absents dans des listes de référence et 35 modèles

n-grammes. Celles-ci sont détaillées dans les trois premiers points de cette section.

Par ailleurs, nous avons également testé diverses variables reposant sur d'autres hypothèses que l'effet de fréquence, à savoir 4 mesures de la diversité lexicale, 13 variables basées sur les caractéristiques des voisins orthographiques et 14 mesures de la longueur des mots. Cela nous donne un total de 170 facteurs lexicaux pris en compte dans cette étude, faisant de celle-ci l'une des comparaisons les plus exhaustives du pouvoir discriminant de ce type de facteurs en lisibilité.

Statistiques de la fréquence lexicale

Une approche simple de l'effet de fréquence, nous l'avons vu, consiste à employer diverses statistiques descriptives pour résumer la distribution des fréquences des mots du texte. Bien que simple à implémenter, cette approche autorise un grand nombre de variantes selon que l'on manipule la catégorie grammaticale des mots considérés, la mesure statistique utilisée et le type d'unités dont est mesurée la fréquence.

Ainsi, en ce qui concerne les catégories de discours, nous avons d'abord défini un ensemble de variables pour l'ensemble des mots des textes. Puis, suivant en ce sens les suggestions de Elley (1969) et Bormuth (1969a), nous n'avons plus considéré que, soit les noms, soit les mots à contenus, ou encore les mots outils, ce qui nous donne quatre ensembles de variables.

D'autre part, nos expérimentations ont également porté sur diverses mesures statistiques. Si la médiane et la moyenne prédominent dans la littérature, elles comportent des points aveugles sans une mesure de dispersion associée. Ainsi, une même moyenne peut correspondre aussi bien à un texte où les mots se situent tous dans une même zone fréquentielle de complexité moyenne qu'à un texte comportant des mots très communs et d'autres très complexes. Il y a fort à parier que ce second texte posera plus de difficultés à la lecture que le premier. Pour chacune des mesures de la fréquence lexicales qui le permettaient, nous avons dès lors calculé la moyenne, la médiane, l'écart interquartile et les valeurs au 75^{e} et 90^{e} percentiles.

Ces mesures ont été obtenues de la manière suivante. À l'aide du TreeTagger (Schmid, 1994), chaque texte est d'abord annoté en catégories de discours. Ces annotations sont ensuite passées au travers d'un filtre afin d'être rendues compatibles avec celles de Lexique3 (New *et al.*, 2004)[27], la base fréquentielle que nous avons sélectionnée pour notre étude. Il s'agit d'un lexique reprenant plus de 50 000 lemmes et plus de 125 000 formes fléchies observées dans un corpus de sous-titres de films d'environ 50 millions de mots (New *et al.*, 2007). Enfin, les fréquences de cette base ont été lissées à l'aide de l'algorithme *Simple*

27. Cette base lexicale est disponible librement à l'adresse : `http://www.lexique.org/`.

Good-Turing décrit par Gale et Sampson (1995), ce qui permet d'attribuer une probabilité aux événements non couverts par ce lexique [28].

Nous sommes ainsi arrivé aux 26 variables suivantes :

MeanFL : Fréquence moyenne des lemmes.

MedianFL : Médiane des fréquences des lemmes.

IQFL : Écart-interquartile des fréquences des lemmes.

75FL : Probabilité équivalant au 75e percentile de la distribution des fréquences des lemmes.

90FL : Probabilité équivalant au 90e percentile de la distribution des fréquences des lemmes.

MeanFLC : Fréquence moyenne du lemme pour les mots de contenu.

MedianFLC : Médiane des fréquences du lemme pour les mots de contenu.

IQFLC : Écart-interquartile des fréquences du lemme pour les mots de contenu.

75FLC : Probabilité équivalant au 75e percentile de la distribution des fréquences du lemme, uniquement pour les mots de contenu.

90FLC : Probabilité équivalant au 90e percentile de la distribution des fréquences du lemme, uniquement pour les mots de contenu.

MeanFLG : Fréquence moyenne du lemme pour les mots grammaticaux.

MedianFLG : Médiane des fréquences du lemme pour les mots grammaticaux.

IQFLG : Écart-interquartile des fréquences du lemme pour les mots grammaticaux.

75FLG : Probabilité équivalant au 75e percentile de la distribution des fréquences du lemme, uniquement pour les mots grammaticaux.

90FLG : Probabilité équivalant au 90e percentile de la distribution des fréquences du lemme, uniquement pour les mots grammaticaux.

MeanFLN : Fréquence moyenne du lemme pour les noms.

MedianFLN : Médiane des fréquences du lemme pour les noms.

IQFLN : Écart-interquartile des fréquences du lemme pour les noms.

75FLN : Probabilité équivalant au 75e percentile de la distribution des fréquences du lemme, uniquement pour les noms.

90FLN : Probabilité équivalant au 90e percentile de la distribution des fréquences du lemme, uniquement pour les noms.

MeanFLV : Fréquence moyenne du lemme pour les verbes.

MedianFLV : Médiane des fréquences du lemme pour les verbes.

MeanFLA : Fréquence moyenne du lemme pour les adjectifs.

MedianFLA : Médiane des fréquences du lemme pour les adjectifs.

MeanFLC-N : Fréquence moyenne du lemme pour les mots de contenu, sauf les noms.

MedianFLC-N : Médiane des fréquences du lemme pour mots de contenu, sauf les noms.

28. La probabilité d'un événement inconnu est -19.6046453987 (en échelle logarithmique à base e), mais les déterminants ordinaux et cardinaux reconnus par le TreeTagger se voient attribuer une log-probabilité de -4.6560646780.

À côté des 26 prédicteurs basés sur les probabilités des lemmes, nous avons également expérimenté les mêmes 26 variables à partir des probabilités des formes fléchies, ce qui représente l'approche la plus commune en lisibilité. Leurs noms sont similaires à ceux décrits ci-dessus, si ce n'est que le **FL** (fréquences des lemmes) est remplacé par **FFF** (fréquence des formes fléchies), ce qui donne, par exemple, **MeanFFF**, **IQFFFC**, etc. Dans un troisième temps, nous avons envisagé les probabilités des formes fléchies désambiguïsées en fonction de leur catégorie grammaticale. Les noms, prenant **FFFD** à la place de **FL**, deviennent dès lors **MeanFFFD**, **MedianFFFD**, **IQFFFDC**, etc.

Pour chacune de ces 78 variables (dont la liste étendue est consultable en annexes), nous avons considéré les occurrences des mots et non les « types », comme il est assez commun de le faire dans la littérature. La raison de ce choix est qu'ayant distingué l'effet de fréquence de l'effet de répétition dans la partie théorique de ce chapitre, il nous a paru nécessaire de conserver cette distinction lors de nos expérimentations. L'effet de répétition est dès lors considéré dans le TTR normalisé (*cf.* Section 3.3.1).

Les absents d'une liste de référence

Une seconde approche de la difficulté lexicale présentée précédemment est le recours à une liste de référence à partir de laquelle est calculé, pour un texte donné, le pourcentage de mots absents de cette référence. Bien que ce type de variables ait été critiqué au profit de celles basées sur les fréquences lexicales (Elley, 1969; Lively et Pressey, 1923), il a néanmoins été largement utilisé en lisibilité. Nous avons dès lors implémenté plusieurs variantes, à la fois dans un but de comparaison, mais aussi pour éclaircir certaines questions qui se posent à leur sujet.

Pour commencer, nous avons employé la liste de Gougenheim *et al.* (1964), un inventaire des 1 063 mots du français considérés comme les plus utiles pour des apprenants de FLE. Il a été élaboré dans les années 50 et est particulièrement intéressant, car il ne comprend pas seulement les termes les plus fréquents de la langue, mais également une série de « mots disponibles ». Sur la base de cette liste, nous avons défini les deux variables suivantes :

PAGoug_1063 : la proportion d'absents [29] de la liste de 1 063 mots de Gougenheim ;

PAGougU_1063 : la proportion de formes différentes (« types ») absentes de la liste de 1 063 mots de Gougenheim.

Par ailleurs, suite aux remarques de Dale et Chall (1948a) ou Harris et Jacobson (1974) sur la relation entre le pouvoir discriminant des variables basées sur le nombre d'absents et la

29. Soulignons que les noms propres, les symboles et les abréviations n'ont pas été considérés dans ce calcul, ainsi que dans ceux qui suivent.

taille de la liste utilisée, nous avons expérimenté des listes de taille différente. Pour ce faire, nous avons repris la liste étendue de Gougenheim, distribuée sur le site de Lexique3 [30], qui comprend 8 875 mots, et dont les 1 063 premiers sont les mêmes que ceux de la liste originale. Sur la base de cet inventaire, nous avons conçu huit listes, comportant respectivement 2 000, 3 000, 4 000, 5 000, 6 000, 7 000, 8 000 et 8 875 mots. À partir de celles-ci, 16 variables supplémentaires ont été définies, suivant ce motif :

PAGoug_X : la proportion d'absents de la liste de X mots de Gougenheim ;

PAGougU_X : la proportion de formes différentes (« types ») absentes de la liste de X mots de Gougenheim.

où X correspond à la taille de l'une des listes précités.

Une faiblesse notable de la liste de Gougenheim est son ancienneté, ce qui a conduit Mesnager (1989) à recommander plutôt l'emploi de l'index de Catach (1985). Celui-ci n'est cependant pas destiné à un public d'apprenants de FLE. C'est pourquoi nous n'avons pas suivi cette suggestion et avons sélectionné à la place une autre liste, qui a été obtenue de la manière suivante. À la fin des manuels de FLE – du moins ceux de niveaux A1 et A2 –, on retrouve régulièrement un index couvrant l'ensemble des mots étudiés dans l'ouvrage. Nous avons ainsi retenu les listes issues des manuels *Alter Ego 1* (Berthet *et al.*, 2006a) et *Alter Ego 2* (Berthet *et al.*, 2006b), qui comprennent respectivement 802 et 1001 mots [31]. Ce vocabulaire récent devrait permettre d'examiner dans quelle mesure l'ancienneté de la liste de Gougenheim influe sur la capacité de prédiction des variables qui en découlent.

Ces deux listes issues d'*Alter Ego* présentent un néanmoins problème : elles négligent les mots grammaticaux, repris par ailleurs dans les synthèses grammaticales. Dès lors, ces termes, dont plusieurs font partie des plus courants de la langue, sont considérés comme absents, ce qui fausse les variables définies sur cette base. Nous avons envisagé deux stratégies pour contourner ce problème : (1) nous avons calculé le pourcentage de mots lexicaux absents ; (2) nous avons intégré à la liste d'*Alter Ego* les 57 mots grammaticaux repris dans les 100 premiers de la liste de Gougenheim *et al.* (1964), ainsi que cinq adverbes très courants (« oui », « non », « aussi », « très » et « pas »). Cela donne 8 variables supplémentaires :

PA_Alterego1a : pourcentage des mots lexicaux absents de la liste de 802 mots d'*Alter Ego 1* ;

PAU_Alterego1a : pourcentage des mots lexicaux différents (« types ») absents de la liste de 802 mots d'*Alter Ego 1* ;

PA_Alterego1b : pourcentage d'absents de la liste de 802 mots d'*Alter Ego 1* + 62 termes ajoutés ;

30. Cette liste est disponible sur la page : `http://www.lexique.org/public/gougenheim.php`.

31. Une autre possibilité pour constituer une liste spécifique au FLE consisterait à utiliser le vocabulaire associé à un niveau particulier dans les référentiels européens de Beacco et Porquier. Nous n'avons toutefois pas envisagé cette perspective, qui mériterait cependant d'être comparée à la liste issue d'*Alter Ego*.

PAU_Alterego1b : pourcentage de formes différentes (« types ») absentes de la liste de 802 mots d'*Alter Ego 1* + 62 termes ajoutés ;

PA_Alterego2a : pourcentage des mots lexicaux absents des 1803 mots d'*Alter Ego* 1 et 2 ;

PAU_Alterego2a : pourcentage des mots lexicaux différents (« types ») absents des 1803 mots d'*Alter Ego* 1 et 2 ;

PA_Alterego2b : pourcentage d'absents de la liste de 1803 mots d'*Alter Ego* 1 et 2 + 62 termes ajoutés ;

PAU_Alterego2b : pourcentage de formes différentes (« types ») absentes de la liste de 1803 mots d'*Alter Ego* 1 et 2 + 62 termes ajoutés ;

Les modèles unigrammes

La vraisemblance textuelle constitue une troisième famille de facteurs qui fait appel aux fréquences lexicales. Cette fois, on ne résume plus la distribution des fréquences via une moyenne ou un quartile, mais on cherche à évaluer la probabilité de la suite de mots du texte à l'aide de modèles n-grammes. Le plus simple d'entre eux, l'unigramme, correspond en réalité à la moyenne géométrique des probabilités de chaque mot. Autrement dit, plutôt que de sommer les probabilités de chaque terme avant de les normaliser, on les multiplie. Dans nos articles précédents (François, 2009a,b), nous avions déjà utilisé avec succès un modèle unigramme (**ML**) comme variable lexicale. Celui-ci attribue, à un texte de longueur n, l'indice de vraisemblance suivant :

$$ML = \prod_{i=1}^{n} P(w_i)^{\frac{100}{n}}$$

La vraisemblance est donc normalisée en fonction de la longueur du document et ramenée à un document de 100 mots, une longueur relativement classique en lisibilité. Ensuite, afin d'éviter les problèmes liés à la représentation et la manipulation informatique de nombres très petits, nous appliquons le logarithme sur l'équation précédente et obtenons la formule suivante :

$$ML = \frac{100}{n} * (\prod_{i=1}^{n} \log P(w_i))$$

Les probabilités utilisées sont les mêmes que pour le calcul des statistiques de la fréquence lexicale et elles ont subi le même lissage. Par conséquent, il a été possible de définir trois variables différentes, selon que l'on prenne en compte la fréquence des lemmes (**ML1**), celle des formes fléchies (**ML2**) ou des formes fléchies désambiguïsées (**ML3**).

Les modèles n-grammes

S'il est relativement simple de développer un modèle unigramme, la conception de modèles d'ordre supérieur est plus complexe, à la fois parce qu'il est difficile d'avoir une bonne couverture lexicale et parce qu'il faut résoudre des questions d'optimisation au niveau de la structure de données utilisée pour stocker les n-grammes et leur fréquence. Confronté à ces problématiques, nous avons (1) sélectionné deux corpus différents au départ desquels sont calculées les fréquences de séquences de 2 à 5 tokens [32] ; (2) appliqué diverses procédures d'optimisation de la structure de stockage des n-grammes.

L'intérêt d'employer deux corpus tient à la possibilité de comparer les effets de leurs caractéristiques – en particulier la taille et le type de documents – sur la capacité prédictive des modèles qui en sont dérivés. Nous avons ainsi utilisé deux corpus qui diffèrent sensiblement par ces deux aspects :

– Les 5-grammes de Google (Michel *et al.*, 2011), qui représentent le plus vaste corpus disponible pour le Français. Une fois les fréquences nettoyées, afin de ne conserver que celles qui concernent des textes publiés entre 2000 et 2008, nous avons obtenu un ensemble de 1 117 140 444 5-grammes. Il faut cependant souligner que la tokenization opérée par les concepteurs de cette ressource très basique : ils considèrent les suites suivantes comme des 5-grammes : « , l ' arbre est » ou « un pique - nique . ». Par ailleurs, la faible taille du vocabulaire de base (181 761 unigrammes) nous a étonné.
– Le second corpus est constitué d'une série d'articles de l'année 2009 du quotidien belge *Le Soir*, pour un total de 5 000 000 de 5-grammes. Dans ce cas, nous avons pu effectuer notre propre découpage en mots et considérer comme une seule unité « pique-nique » ou « l' ». La taille du vocabulaire de base atteint ici les 117 115 unigrammes.

Nous avons ensuite stocké chaque n-gramme et sa fréquence dans une structure appelée PATRICIA tree (Morrison, 1968). Cette structure de données autorise la compression des n-grammes qui partagent un préfixe commun et des noeuds ne possédant qu'un seul noeud-fils, ce qui permet d'effectuer des requêtes en temps constant. Nous avons encore amélioré le taux de compression en indexant le vocabulaire dans une table séparée où chaque mot est associé à un entier, qui est nettement moins consommateur en termes de mémoire.

Sur la base de cette structure de données, nous avons défini 12 variables pour les modèles d'ordre 2 et 3 [33]. La liste qui suit les présente de manière condensée : X correspond à la base de n-grammes employées (« LS » pour *Le Soir* et « G » pour *Google*) ; quant à Y, il

32. Il ne nous semble pas ici adapté de parler de « mots » pour référer aux éléments constitutifs des n-grammes, car il peut s'agir de signes de ponctuation, d'une partie de mots composés, etc. Nous utilisons dès lors le terme *token*, emprunté à l'anglais.

33. Les modèles d'ordre supérieurs n'ont pas été évalué de manière systématique, car il est vite apparu que la présence de nombreuses données inconnues du modèle rendait les prédicteurs inefficaces.

représente l'ordre n du modèle n-gramme.

normTLProb-X-Y[34] **:** la log-vraisemblance de chaque texte, normalisée tel que proposé par Kate *et al.* (2010), et exprimée comme suit :

$$\text{normTLProb} = \frac{1}{m}\sum_{i=1}^{m}\log P(w_i|h) \tag{3.13}$$

où $P(w_i|h)$ correspond à la probabilité conditionnée à l'historique h, limité au $n-1$ mots précédents, et m représente le nombre de mots dans le texte à analyser. Il s'agit du modèle n-gramme classique ;

MPCond-X-Y : la moyenne des probabilités conditionnelles utilisées dans le modèle précédent. La différence avec l'approche précédente est donc que l'on additionne les probabilités plutôt que de les multiplier ;

MedPCond-X-Y : la médiane des probabilités conditionnelles ;

Par ailleurs, nous avons aussi défini 20 variables reposant, non pas sur les probabilités conditionnelles ($P(w_i|h)$), mais sur les probabilités simples des suites $P(w_i \cap h)$. Leurs noms respectent également le schéma précédent.

meanNGProb-X-Y : la moyenne arithmétique des probabilités simples des suites de taille X présentes dans le texte ;

gmeanNGProb-X-Y : la moyenne géométrique de ces mêmes probabilités (équivalent de **normTLProbX**, en ce sens que les probabilités sont ici multipliées) ;

medianNGProb-X-Y : la médiane de ces probabilités ;

NGprob75-X-Y : ainsi que le 75^{e} percentile ;

NGprob90-X-Y : et le 90^{e} percentile.

Afin d'éviter le problème bien connu surgissant lorsqu'une suite de mots n'a pas été observée dans la base de n-grammes, nous avons appliqué à ces différentes probabilités un lissage par interpolation linéaire. Bien qu'il ne s'agisse pas de la technique de lissage la plus performante d'entre toutes, elle reste tout à fait dans la moyenne des algorithmes existants. De plus, nous avons effectué ce choix, car il est possible de l'appliquer à la volée, sans qu'il soit nécessaire d'effectuer une série de calculs préalables requérant un nombre extrêmement élevé de requêtes sur la base de données, comme c'est le cas pour les approches basées sur le « discouting », c'est-à-dire une technique où une partie de la masse probabiliste des événements observées est décomptée et réservée aux nouveaux événements.

L'interpolation linéaire est basée sur l'idée que la probabilité d'une suite d'ordre n peut être est approximée comme la somme de sa propre probabilité et de celles des suites d'ordre inférieur, ainsi que le décrit la formule récursive suivante (Chen et Goodman, 1999, 312) :

$$p_{interp}(w_i \mid W_{i-n+1}^{i-1}) = \lambda_{W_{i-n+1}^{i-1}} p_{ML} * (w_i \mid W_{i-n+1}^{i-1}) + (1 - \lambda_{W_{i-n+1}^{i-1}}) * p_{interp}(w_i \mid W_{i-n+2}^{i-1}) \tag{3.14}$$

où $p(w_i \mid W_{i-n+1}^{i-1})$ représente la probabilité du mot w_i, interpolée (p_{interp}) ou issue directement du modèle (p_{ML}), étant donné son historique W_{i-n+1}^{i-1}. Comme cette formule est récursive, elle doit s'arrêter à un moment donné, généralement l'ordre 0, pour lequel est défini une distribution uniforme obtenue à partir de la taille du vocabulaire V :

$$p_{unif}(w_i) = \frac{1}{\mid V \mid} \tag{3.15}$$

Notons que nous n'avons pas entraîné des valeurs différentes pour $\lambda_{W_{i-n+1}^{i-1}}$, ce qui ne s'avère généralement pas heureux (Chen et Goodman, 1999, 312). À la place, nous avons utilisé une heuristique pour attribuer la valeur $0,9$ à λ.

Les mesures de la diversité lexicale

À côté des variables basées sur la fréquence objective, nous avons proposé quatre paramétrisations de l'effet de répétition, qui recourent toutes au TTR. Dans la section 3.2.2, nous avions insisté sur les limitations du TTR, eu égard à la difficulté de le normaliser. Toutefois, nous avons malgré tout testé celui-ci, sous quatre formes différentes :

TTR_W : le TTR classique, non normalisé et calculé sur la base des formes fléchies.

TTR_L : le TTR non normalisé, mais calculé à partir des lemmes. Cette approche inhabituelle nous semble intéressante à expérimenter à partir du moment où les lecteurs de FLE sont généralement capables d'analyser correctement les formes fléchies et qu'il nous apparaît qu'une forme très proche d'une autre devrait jouer un rôle facilitateur.

NormTTR_W : une version normalisée du TTR au départ des mots. La technique de normalisation employée est inspirée du manuel de WordsmithTool (*cf. op. cit.*). Elle découpe les textes en sections de n mots sur lesquels est calculé un TTR et effectue la moyenne des résultats obtenus pour chaque section. La valeur de n peut évidemment impacter l'efficacité de la variable.

NormTTR_L : une version normalisée du TTR qui utilise la même procédure que **NormTTR_W**, mais en utilisant les lemmes.

Il faut insister sur le fait que la version normalisée du TTR repose sur un méta-paramètre : n. La valeur de ce dernier peut évidemment influencer l'efficacité des variables qui en découlent. Trop large, il atténue l'effet de la normalisation, puisque de nombreux textes pour débutants seront de taille inférieure ; trop petit, il s'éloigne de l'objectif du TTR et de la réalité psychologique de l'effet de répétition. Dans ce cas, en effet, le modèle postule qu'à chaque nouveau fragment de n mots, le lecteur oublie les mots qu'il a rencontré auparavant et que l'effet de répétition ne peut dépasser ces frontières étroites. Afin de sélectionner une valeur de n optimale, nous allons explorer le continuum compris entre 100 (une taille de fragment textuel habituel en lisibilité) à 1 000, qui représente le choix par défaut dans WordsmithTool.

Le voisinage orthographique

Notre revue des connaissances théoriques sur le voisinage orthographique dans le cadre bilingue (*cf.* Section 3.2.2) a insisté sur le fait qu'aussi bien les voisins orthographiques de la L1 que ceux de la L2, influencent la vitesse de reconnaissance d'un mot donné. Toutefois, la prise en compte de la situation des bilingues est relativement complexe et spécifique à une paire de langues précise. Dans le cadre de cette thèse, dédiée à une formule de lisibilité indépendante de la L1 des lecteurs, nous avons simplifié ce tableau, en considérant les lecteurs en L2 comme des lecteurs de L1. Autrement dit, seuls les voisins propres à la L2 sont considérés.

Dans la partie théorique, nous avons insisté sur les résultats divergents entre le français et l'anglais. Dans la langue de Corneille, la présence d'un voisin plus fréquent inhiberait la reconnaissance des mots, tandis qu'en anglais, on a observé un effet facilitateur lorsque la densité du voisinage augmente. Comme les résultats des différentes études demandent encore à être confirmés, nous avons toutefois envisagé ces deux hypothèses pour le français.

Nous nous sommes basé sur la liste de mots et de leurs voisins distribuée dans le cadre du projet Lexique3 [35]. Celle-ci comprend 128 919 formes fléchies pour lesquelles il est précisé à chaque fois le nombre de voisins, la liste de ces voisins et leur probabilité, calculée à partir du corpus Frantext [36]. Il nous a paru que la taille de ce vocabulaire devrait couvrir la plupart des formes avec lesquelles un lecteur en L2 est susceptible d'entrer en contact. Les mots absents de la liste n'ont simplement pas été considérés et nous avons calculé pour chaque texte à analyser le pourcentage de mots non pris en compte. Nous avons également négligé les noms propres, qui ne sont pas repris dans le pourcentage d'absents de la liste. Par ailleurs, les probabilités des 66 712 voisins se sont vues appliquer un lissage à l'aide de l'algorithme *Simple Good-Turing* (Gale et Sampson, 1995). Cette opération a permis d'attribuer une probabilité aux voisins listés, mais qui n'ont pas été observés dans le corpus Frantext.

Une fois cette série de pré-traitements terminée, nous avons calculé, pour chaque mot des textes à analyser, les 4 valeurs suivantes :

nbNeigh : le nombre de voisins ;

neighMoreFreq : la présence d'un voisin plus fréquent (prend comme valeur 0 ou 1) ;

nbNeighMoreFreq : le nombre de voisins plus fréquents ;

freqCumNeigh : les fréquences cumulées de tous les voisins.

Comme précédemment, ces valeurs forment, pour un texte donné, une distribution dont il est possible de calculer diverses statistiques. À l'exception de **neighMoreFreq** où seule

35. Elle est disponible gratuitement à l'adresse suivante : `http://www.lexique.org/telAutresBases.php`.

36. Ce corpus est consultable à l'adresse `http://www.frantext.fr/`.

la moyenne a été considérée, pour les trois autres distributions, nous avons envisagé la moyenne, la médiane, et les 75e et 90e percentiles. Cela donne au final 13 variables basées sur les voisins.

La longueur des mots

Enfin, il n'aurait pas été possible de se passer des variables lexicales les plus populaires : celles basées sur la longueur des mots. À ce niveau, nous avons repris la plupart des variables utilisées dans la littérature et présentées dans la première partie de ce chapitre. Les 14 prédicteurs suivants ont ainsi été implémentés :

NLM : Le nombre moyen de lettres par mots ;

NMLM : Le nombre médian de lettres par mots ;

PM+4 : La proportion de mots de plus de 4 lettres ;

PM+5 : La proportion de mots de plus de 5 lettres ;

PM+6 : La proportion de mots de plus de 6 lettres ;

PM+7 : La proportion de mots de plus de 7 lettres ;

PM+8 : La proportion de mots de plus de 8 lettres ;

PM+9 : La proportion de mots de plus de 9 lettres ;

PM+10 : La proportion de mots de plus de 10 lettres ;

PM+11 : La proportion de mots de plus de 11 lettres ;

PM+12 : La proportion de mots de plus de 12 lettres. Nous n'avons pas expérimenté une valeur supérieure à 12, car le pourcentage devient très réduit ;

NL3Q : Le nombre de lettres correspondant au troisième quartile (ou 75e percentile) de la distribution de la longueur des mots du texte ;

NL80P : Le nombre de lettres correspondant au 80e percentile ;

NL90P : Le nombre de lettres correspondant au 90e percentile.

Comme le lecteur averti l'aura noté, les trois derniers facteurs cités représentent l'opération statistique inverse des mesures PM+x sur la distribution de la longueur des mots du texte. Ainsi, NL3Q, par exemple, équivaut à x, la valeur en abscisse où la fonction de répartition vaut $F(x) = 0,75$, tandis que les mesures PM+x s'expriment comme $100 * (1 - P(X \leq x))$, où $P(X \leq x)$ est équivalent à $F(x)$. Ces différents facteurs visent à prendre en compte d'autres aspects de la distribution que la simple moyenne (NLM), et ce afin de distinguer des textes dont les variances diffèrent nettement.

Pour obtenir ces différentes mesures, nous sommes partis de la découpe des textes en mots réalisée par le TreeTagger (Schmid, 1994). Nous avons ensuite dénombré le nombre de caractères de chaque mot, sans tenir compte des apostrophes, des tirets ou du symbole de

degré. Nous avons ainsi obtenu une liste avec le nombre de lettres de chaque mot du texte qui a été traitée avec le module **stats** de python afin d'en dériver nos 14 variables d'intérêt.

3.3.2 Les variables syntaxiques

Les indices de la difficulté syntaxique représentent une seconde source d'information pour prédire la difficulté des textes. De la même façon que l'on a pensé que les mesures basées sur la fréquence des mots seraient plus efficaces que celles reposant sur le nombre d'absents d'une liste de référence, plusieurs chercheurs espéraient que les variables basées sur la complexité syntaxique seraient plus prédictives que celles basées simplement sur la longueur phrastique. Cependant, ni Bormuth (1966), ni MacGinitie et Tretiak (1971), ni d'autres encore, ne sont parvenus à entériner cette hypothèse, obtenant au contraire des corrélations inférieures. Par ailleurs, les mesures de la complexité phrastique nécessitent l'intervention d'un parseur syntaxique, dont les résultats restent encore peu fiables pour le français.

Pour ces deux raisons, nous n'avons pas considéré les facteurs de la complexité syntaxique, quand bien même ceux-ci sont plus fondés théoriquement. À la place, nous avons envisagé les trois types de prédicteurs suivants : ceux qui informent sur la longueur des phrases des textes (12) ; ceux qui décrivent des ratios entre catégories de discours (156) et ceux qui prennent en compte la difficulté des verbes (23). Cela nous fait donc un total de 191 indices pour le niveau syntaxique.

La longueur des unités syntaxiques

Comme nous y avons déjà insisté, mesurer automatiquement la longueur d'une phrase est relativement simple. Une fois que le texte à analyser a été annoté par TreeTagger (Schmid, 1994), nous comptons simplement le nombre de tokens (excepté les signes de ponctuation) par phrase (le découpage est aussi réalisé par TreeTagger), obtenant ainsi une distribution pour ce texte. À partir de celle-ci, nous avons défini les 6 variables suivantes :

meanNWS : le nombre moyen de mots par phrase ;

medNWS : le nombre médian de mots par phrase ;

NWS_75 : la taille de la phrase correspondant au 75^{e} percentile de la distribution ;

NWS_90 : la taille de la phrase correspondant au 90^{e} percentile de la distribution ;

LenSentDolch : correspond à **NWS_90** - **medNWS**, une mesure définie par Dolch (1948) ;

LenSentDaoust : le pourcentage de phrases de plus de 30 mots, utilisé par Daoust *et al.* (1996).

Une question qui se pose dans la paramétrisation de ces variables est la manière dont on définit une phrase. Nous avons opté pour deux délimitations différentes. La première se limite

aux points de ponctuation forts : « ! », « . » et « ? » et correspond à une définition conservatrice de cette unité linguistique. La seconde considère également les trois points de suspension, lorsque ceux-ci sont suivis d'une majuscule ou d'un tiret, ce qui est indicatif d'une prise de parole dans un dialogue. Nous obtenons dès lors 6 variables supplémentaires portant les noms de celles décrites ci-dessus, auxquels est adjoint un 2 à la fin (ex. : **medNWS2**).

Les catégories du discours

Pour cette famille de variables, nous nous sommes basé sur l'étude de Bormuth (1966), qui est l'une des plus complètes sur la question. Elle montre l'intérêt de ne pas considérer uniquement les odds [37] d'une catégorie contre toutes les autres, mais aussi de calculer les odds entre deux ou plusieurs catégories. Nous avons repris cette façon d'aborder les variables catégorielles.

Pour ce faire, nous sommes partis des textes annotés par le TreeTagger (Schmid, 1994), lequel propose un jeu de 33 catégories [38] que nous avons simplifié pour obtenir douze classes plus générales. L'énumération suivante détaille comment 10 de ces classes ont été obtenues sur la base des étiquettes de TreeTagger :

ADV : ADV

NOM : NOM

ART : ART, PRP :det, DET :ART

PRO : PRO, PRO :DEM, PRO :IND, PRO :PER, PRO :POS, PRO :REL

NAM : NAM

NOM+NAM : NOM, NAM

PRE : PRP

VER : VER :cond, VER :futu, VER :impe, VER :impf, VER :infi, VER :pper, VER :ppre, VER :pres, VER :simp, VER :subi, VER :subp,

ADJ : ADJ, DET :POS, NUM

CON : KON

Nous n'avons pas considéré les abréviations (ABR), la ponctuation (PUN, PUN :cit et SENT), les symboles (SYM), ni les interjections (INT). Par contre, nous avons ajouté deux classes aux dix susmentionnées : **GRAM**, qui reprend les mots grammaticaux et **LEX**, les mots lexicaux.

Enfin, sur la base de ces douze classes, nous avons calculé pour chacune d'elles : (1) ses odds par rapport à chaque autre catégorie ; (2) son odd contre toutes les autres catégories et

37. Pour deux catégories, a et b, il s'agit de $\frac{p(a)}{p(b)}$, c'est-à-dire le ratio entre la probabilité de a et celle de b.

38. Ce jeu a été développé par Achim Stein en 2003 et est disponible à l'adresse suivante : `http://www.ims.uni-stuttgart.de/~schmid/french-tagset.html`.

(3) sa proportion par rapport aux mots du texte. Cela nous donne 13 variables pour chacune des 12 catégories de discours retenues, c'est-à-dire 156 variables catégorielles. Elles sont appelées d'après le ratio qui est calculé à chaque fois. Par exemple, **GRAM/ADV** représente le rapport des mots grammaticaux aux adverbes, tandis que **VER/ALL** est la proportion de verbes dans le texte.

Les verbes

La troisième dimension syntaxique prise en considération dans cette étude concerne les aspects verbaux. Comme nous l'avons expliqué précédemment dans ce chapitre, il n'existe guère de modèles à suivre quant leur paramétrisation, si ce n'est Daoust *et al.* (1996) qui suggèrent un ratio des verbes fléchis sur le nombre de mots du texte et Heilman *et al.* (2007), qui repèrent quelques formes verbales. Dans ce dernier cas, les auteurs ne précisent pas la forme que prennent leurs variables. Nous supposons qu'il s'agit de prédicteurs binaires, ainsi que nous l'avons nous-même implémenté précédemment (François, 2009a). Ceux-ci attestent simplement de la présence de certains temps dans le texte à évaluer, mais sont particulièrement adaptés au contexte FLE. En effet, comme l'apprentissage grammatical par étapes va de pair avec l'introduction progressive des temps verbaux, il est clair qu'il existe des différences notables entre les profils de textes A2 et C1.

Afin de repérer les temps verbaux dans les textes, nous avons utilisé une fois de plus le TreeTagger. Celui-ci est capable de détecter automatiquement 11 « temps » : le conditionnel (**Cond**), le futur (**Futur**), l'impératif (**Imperatif**), l'imparfait (**Impf**), l'infinitif (**Infi**), le participe passé (**PPasse**), le participe présent (**PPres**), le présent de l'indicatif (**Pres**), le passé simple (**PasseSim**), le subjonctif imparfait (**Subi**) et le subjonctif présent (**Subp**). Nous avons dès lors défini 11 variables binaires correspondant à ces étiquettes, même s'il est vrai que cette classification n'est pas optimale. Certains temps sont en effet réunis (ex. futur simple et futur antérieur), alors que leur complexité diffère très certainement. Néanmoins, la majorité des temps principaux sont distingués, ce qui devrait suffire à discriminer les textes pour débutants des textes plus avancés.

Si ces indices binaires ont fait la preuve de leur efficacité, on peut légitimement douter qu'ils suffisent à modéliser des différences plus fines. Ainsi, ils ne nous apprennent rien lorsque deux textes partagent les mêmes temps verbaux, mais que ceux-ci sont répartis de façon différente. En effet, on peut raisonnablement penser qu'un extrait comportant une majorité de verbes au passé simple sera plus ardu qu'un autre où le présent de l'indicatif domine. C'est pourquoi nous avons implémenté une version continue de ces mêmes variables, qui correspond à la proportion de verbes conjugués à un des 11 « temps » susmentionnés sur le

total des verbes présents dans le texte. Cela nous donne 11 prédicteurs supplémentaires [39]. Par ailleurs, nous avons également implémenté le ratio de Daoust *et al.* (1996) (**ratioDaou**) à des fins de comparaison.

3.3.3 Les aspects sémantiques

Pendant longtemps, la majorité des formules de lisibilité se sont limitées aux deux niveaux de facteurs que nous venons de décrire. L'une des explications à cette situation est la complexité des modèles mis en oeuvre, qui rendait longue et pénible leur application manuelle. D'autant que, comme les structuro-cognitivistes avaient dû le reconnaître, la prise en compte des aspects sémantiques ne semblait guère apporter d'information supplémentaire à des variables plus conventionnelles. Récemment, l'application de techniques de TAL pour automatiser la paramétrisation des variables sémantiques a laissé espérer qu'il soit désormais possible de trancher ce délicat dilemme entre des modèles vraisemblables, mais complexes et des approches pratiques, mais superficielles.

Plusieurs travaux récents (Crossley *et al.*, 2007; Feng *et al.*, 2009, 2010; Graesser *et al.*, 2004; Pitler et Nenkova, 2008) se sont ainsi concentrés sur le développement de nouvelles variables sémantiques ou l'implémentation de facteurs connus, mais précédemment employés via une annotation manuelle. Cette partie des recherches modernes se situe au coeur de notre étude, puisque celle-ci vise en notamment à évaluer les apports du TAL à la lisibilité du FLE. C'est pourquoi nous avons implémenté au moins une variable pour chacune des trois grandes familles de prédicteurs sémantiques : le niveau d'abstraction du vocabulaire, la densité des idées et le taux de cohérence des textes, via une approche par modèle à vecteurs de mots. Ces trois aspects sont détaillés dans le reste de cette section.

Taux d'abstraction et de personnalisation du texte

À côté des modèles TAL complexes envisagés dans la suite de cette section, nous avons sélectionné deux ensembles de prédicteurs relativement simples qui visent à appréhender le niveau de concrétude des textes, mais également le taux de personnalisation. Ceux-ci ont déjà été discutés précédemment (*cf.* Section 3.2.4) et n'ont pas posé de difficultés particulières, si ce n'est en ce qui concerne la technique de normalisation pour les classes de pronoms.

En effet, il est possible de calculer la proportion de pronoms par rapport au nombre de mots du texte ou relativement aux autres classes. Dale et Tyler (1934) ne normalisaient pas direc-

39. Les noms sont similaires à ceux des variables binaires, si ce n'est que le caractère continu est indiqué par la chaîne « _C » en fin de nom.

tement, vu qu'ils travaillaient sur des fragments de taille similaire. Leur décompte du nombre de pronoms revient cependant à une normalisation en fonction de la longueur du texte. En ce qui concerne Henry (1975), le manuscrit reste peu explicite, même s'il nous semble que c'est la même démarche qui a été adoptée. Dans notre étude, nous avons cependant envisagé les deux possibilités, ce qui nous donne donc les 13 variables suivantes :

PP1 : Pourcentage de pronoms personnels de la 1re personne (je, me, moi, nous) sur le nombre de mots du texte ;

PropP1 : Proportion de pron. pers. de la 1re personne par rapport à l'ensemble des pron. pers. ;

PP2 : Pourcentage de pron. pers. de la 2e personne (tu, te, toi, vous) sur le nombre de mots du texte ;

PropP2 : Proportion de pron. pers. de la 2re personne par rapport à l'ensemble des pron. pers. ;

PP3 : Pourcentage de pron. pers. de la 3e personne (il, elle, le, la, lui, ils, elles, les, leur, eux) sur le nombre de mots du texte ;

PropP3 : Proportion de pron. pers. de la 3re personne par rapport à l'ensemble des pron. pers. ;

PP2_S : Pourcentage de pron. pers. de la 2e personne au singulier (tu, te, toi) sur le nombre de mots du texte ;

PropP2_S : Proportion de pron. pers. de la 2e personne au singulier par rapport à l'ensemble des pron. pers. ;

PP2_P : Pourcentage de pron. pers. de la 2e personne au pluriel (vous) sur le nombre de mots du texte ;

PropP2_P : Proportion de pron. pers. de la 2e personne au pluriel par rapport à l'ensemble des pron. pers. ;

PP1P2 : Pourcentage de pron. pers. de la 1re et la 2e personne sur le nombre de mots du texte [40] ;

PropP1P2 : Proportion de pron. pers. de la 1re et la 2e personne par rapport à l'ensemble des pron. pers. ;

ADJ/VER : Ratio du nombre d'adjectifs sur le nombre de verbes (déjà défini dans la section sur les catégories de discours), qui d'après Broder (1940, cité par Bloomer, 1959, 269, est indicatif du niveau d'abstraction.

Ces différentes mesures vont permettre de comparer les différentes positions théoriques concernant la meilleure façon d'approcher le taux de personnalisation, dont nous avons discuté dans la partie théorique de ce chapitre.

La densité des idées

Afin de mesurer la densité des idées présentes dans les textes de notre corpus, nous avons utilisé le logiciel libre *Densidées* (Lee *et al.*, 2010) [41]. Celui-ci permet de calculer, pour le français, la densité propositionnelle telle que Kintsch, ainsi que Turner et Greene l'ont définie (*cf.* Section 3.2.4).

Le logiciel s'inspire d'un programme similaire pour l'anglais (Brown *et al.*, 2008), à même d'approximer cette densité propositionnelle à l'aide d'un simple taggeur. En effet, Snowdon

40. Cette variable, très similaire à **PPD** définie ci-dessous (*cf.* Section 3.3.4), s'en différencie en ce qu'elle ne prend pas en compte les pronoms possessifs.

41. Ce logiciel est accessible gratuitement à l'adresse `http://code.google.com/p/densidees/`.

et al. (1996) avaient déjà montré, dans une étude sur les relations entre la perte des capacités cognitives en fin de vie et les productions écrites en début de vie adulte, que la densité des propositions peut être estimée à partir du nombre de verbes, d'adjectifs, d'adverbes, de prépositions et de conjonctions, divisé par le nombre de mots. Brown *et al.* (2008) suggèrent d'affiner cette approche à l'aide de 37 règles qui prennent en compte certaines exceptions, c'est-à-dire des termes appartenant aux catégories grammaticales précitées, mais ne représentant pas une proposition (ex. *to*, suivi d'un verbe ; *or* précédé de *either*, etc.). Ce système, appelé CPIDR 3, se révèle capable de reproduire la majorité des exemples de Turner et Greene (1977, cités par Brown *et al.*, 2008, 543), et est corrélé à $0,97$ avec la moyenne de deux jugements humains.

Lee *et al.* (2010) reprennent donc cette idée de partir des verbes (prédicateurs), adjectifs et adverbes (modifieurs) et des prépositions et conjonctions (connecteurs) et adaptent le compte des propositions en fonction de règles, qui sont cette fois au nombre de 35. Le logiciel recourt à TreeTagger pour l'étiquetage morphosyntaxique, lequel ne semble pas trop perturber le calcul de la mesure, puisque les auteurs rapportent que les faux négatifs (prédicats non étiquetés comme tels) et les faux positifs (non-prédicats étiquetés comme des prédicats) ont tendance à s'annuler. *Densidées* obtient ainsi lui aussi un coefficient de corrélation élevé ($r = 0,972$) lorsque ses résultats sont comparés à ceux obtenus manuellement.

En nous aidant de ce logiciel, dont nous avons légèrement modifié le code afin d'adapter les entrées et sorties du programme à nos besoins, nous avons défini la variable suivante :

ConcDens : Le nombre de propositions totales dans le texte divisé par le nombre de mots.

Ce résultat diffère donc de la sortie habituelle du logiciel, qui le multiplie par 10, afin d'obtenir une mesure de densité des idées pour 10 mots.

Le taux de cohérence mesuré via LSA

La section 3.2.5 a détaillé comment la cohérence locale de textes pouvait être approximée à l'aide de diverses mesures basées sur une représentation vectorielle des textes. Parmi celles-ci, nous avons opté pour les suivantes : le cosinus moyen entre phrases adjacentes dans un simple espace vectoriel pondéré (**coRefLocal1**)[42] et dans un espace sémantique obtenu par LSA (**avLocalLsa**).

La première étape à réaliser pour calculer ces deux indices consiste à définir l'espace sémantique voulu à partir d'un ensemble de documents d'apprentissage représentés sous la forme de vecteurs de fréquence. Pour ce faire, nous avons sélectionné un sous-ensemble des textes

42. Cette mesure correspond à la formule de la *Co-reference cohesion local-1* de Graesser *et al.* (2004), si ce n'est qu'elle est définie sur les mots et non les noms.

de notre corpus selon une procédure décrite par après (*cf.* Chapitre 5). Chaque texte a subi, à l'aide de TreeTagger, une série de prétraitements visant à le découper en une suite de mots, laquelle a ensuite été transformée en un simple « sac de mots ». Cet ensemble de mots est ensuite passé au travers d'un filtre – une liste de 126 formes fléchies – destiné à supprimer les termes les plus courants, dont l'apport informatif au modèle serait négligeable. Au terme de ces prétraitements, tous les termes rencontrés dans le corpus ont été indexés au sein d'un dictionnaire où chaque mot est associé à un indice. Outre qu'il réduit l'empreinte mémorielle de la matrice des termes-documents, ce dictionnaire sert à définir les dimensions de l'espace sémantique dans lequel sont projetés les documents.

Cependant, avant d'être projeté dans cet espace sémantique, chaque vecteur est transformé à l'aide d'une fonction de pondération qui évalue la spécificité de chacun de ses termes par rapport au corpus, afin de donner plus de poids aux termes répétés, mais qui sont répartis dans un nombre restreint de documents. Dans leur article, Foltz *et al.* (1998) avaient employé la log-entropie (Dumais, 1991, cité par Bestgen, 2011, 108) dans ce but, mais nous avons préféré utiliser le tf-idf. Celui-ci est aujourd'hui largement répandu dans le domaine de la recherche d'information et il s'agissait également du seul disponible dans l'environnement de développement **gensim** (Řehůřek et Sojka, 2010) [43] que nous avons utilisé pour la création de l'espace vectoriel et pour effectuer ses transformations.

Comme son nom l'indique, le *tf-idf* est la combinaison de deux indices : le $tf_{i,j}$ ou fréquence relative du terme i dans le document j et l'idf_i (Sparck Jones, 1972) ou fréquence inverse de document, qui mesure l'importance du terme dans l'ensemble du corpus à l'aide de la formule suivante (Jurafsky et Martin, 2009, 805) :

$$idf_i = log(\frac{N}{n_i}) \tag{3.16}$$

Celle-ci indique que l'idf_i d'un terme se calcule comme le logarithme du nombre N total de documents divisé par le nombre n_i de documents où apparaît au moins une fois le terme i. Pour obtenir le *tf-idf*, on multiplie simplement ces deux indices.

Une fois cette première transformation effectuée, il est possible d'y ajouter une seconde : la LSA. Cette technique développée par Deerwester *et al.* (1990) consiste à réduire le rang de la matrice des termes-documents (X) à l'aide d'une procédure mathématique complexe : la SVD. Cette opération décompose X, la matrice des termes-documents sous la forme du produit de trois autres matrices : $U\Sigma V^T$, où U est la matrice des termes, V^T est la matrice des documents et Σ contient un ensemble de k valeurs singulières sur sa diagonale (*cf.* Figure 3.1). Ce sont ces dernières qui sont manipulées dans la LSA, dont l'objectif est de

43. Cet outil est disponible à l'adresse `http://nlp.fi.muni.cz/projekty/gensim/`.

réduire l'espace sémantique des mots définis par le dictionnaire à un espace sémantique plus petit défini à partir des k meilleures valeurs singulières [44]. On obtient alors les matrices U_k, Σ_k et V_k^T de rang k.

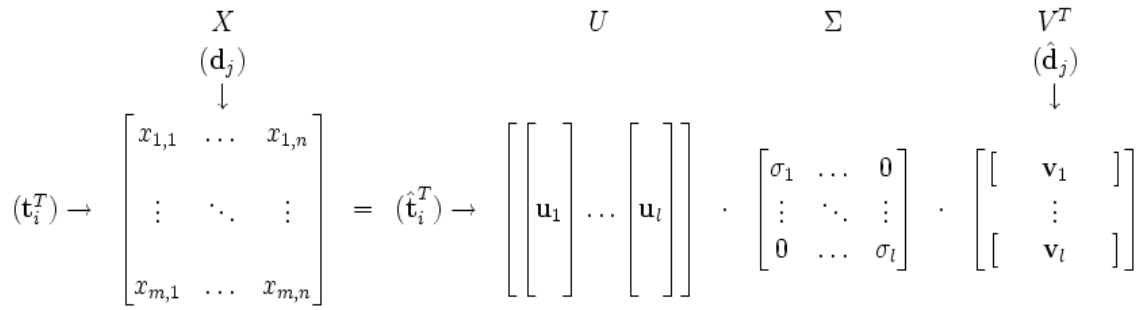


Figure 3.1 – Image tirée de Wikipédia : http ://en.wikipedia.org/wiki/Latent_semantic_analysis

L'avantage de cette simplification n'est pas tant de réduire l'espace, comme c'est le cas dans la technique proche de l'analyse factorielle, mais de faire ressortir la structure sémantique cachée derrière les mots précis employés dans les documents. Par ce moyen, la LSA permet de détecter des similarités entre des textes qui partagent beaucoup de synonymes ou de thématiques communes, mais peu de termes semblables. Il s'agit là de l'avantage principal par rapport à une simple comparaison entre des vecteurs de mots transformés par *tf-idf*.

Une fois l'espace vectoriel constitué, que ce soit par simple transformation par *tf-idf* ou à l'aide d'une LSA, il reste à y projeter les textes dont il faut évaluer la lisibilité. Nous leur avons appliqué les mêmes prétraitements qu'aux documents constitutifs de l'espace sémantique, si ce n'est que nous avons travaillé au niveau des phrases et non plus des documents. La découpe en phrases a été effectuée à l'aide de TreeTagger et chacune d'elles est dès lors représentée sous la forme d'un vecteur numérique q_j. Ce vecteur, situé dans l'espace des mots du dictionnaire, doit lui aussi subir les mêmes transformations que l'espace dans lequel on désire le projeter. Pour ce qui est du *tf-idf*, son calcul est aisé vu que **gensim** stocke les valeurs idf_i pour chaque mot du dictionnaire.

Le cas de la LSA est un peu plus complexe. On cherche à projeter le vecteur de mots q_j dans l'espace de mots sémantique réduit afin d'obtenir le vecteur $\hat{q}_j$. On peut dès lors effectuer pour ce vecteur la même décomposition en valeurs singulières que précédemment pour X, ce qui nous donne :

$$q_j = U_k \Sigma_k \hat{q}_j \tag{3.17}$$

Comme on cherche à obtenir $\hat{q}_j$ et non q_j, il suffit de transformer la relation ainsi (Deerwester

44. Pour obtenir plus de détails sur les aspects mathématiques de la LSA, le lecteur peut se reporter à la synthèse de Manning *et al.* (2008) ou à l'article original de Deerwester *et al.* (1990), qui reste tout à fait abordable.

et al., 1990, 399) :

$$\hat{q}_j = q_j^T U_k \Sigma_k^{-1} \tag{3.18}$$

Cette opération est évidemment disponible dans **gensim**, qui génère un vecteur $\hat{q}_j$ pour chaque phrase j du texte à évaluer. Reste alors à calculer le cosinus[45] entre les vecteurs réduits de deux phrases adjacentes avant d'en faire la moyenne (*cf.* Équation 3.9).

Suite à cette implémentation, nous avons pu définir deux versions de **coRefLocal1** et de **avLocalLsa**, selon que l'on prenne en compte les formes fléchies ou les lemmes. Soulignons enfin que le choix du paramètre k est crucial pour la LSA. Deerwester *et al.* (1990) estiment que 50 à 100 dimensions suffisent. Toutefois, ce nombre dépend également du nombre de documents utilisés pour créer l'espace. Pour un large corpus, Landauer (2004, cité par Bestgen, 2011, 109) recommande plutôt 300 facteurs orthogonaux. Quant à Manning *et al.* (2008, 382), ils rapportent que, dans différents contextes applicatifs, le rappel tend à augmenter tandis que la valeur de k diminue. Par conséquent, nous avons dû effectuer des expériences par validation extrinsèque afin d'évaluer la valeur de k adaptée à nos données. Celles-ci sont détaillées dans le Chapitre 5.

3.3.4 Variables spécifiques aux langues étrangères

Enfin, la dernière famille de variables que nous explorerons dans cette thèse concerne les aspects propres à la pratique de la lecture en FLE. Ne disposant pas d'un corpus adéquat, nous n'avons pas pu implémenter des modèles spécifiques afin de traiter les congénères. Par contre, nous avons longuement exploré une autre problématique des langues étrangères : les unités polylexicales[46], sur lesquelles ont été basés 22 indices de lisibilité. Ceux-ci visent à évaluer l'apport de la notion d'unités polylexicales à la lisibilité des L2. Par ailleurs, nous avons également défini 5 facteurs traitant du type de texte.

Les unités polylexicales

La paramétrisation des variables reposant sur les UPs nous a demandé davantage d'efforts de programmation, puisqu'il a fallu mettre au point un extracteur automatique d'UPs capable de traiter des textes courts. En effet, si les réalisations actuelles dans le domaine donnent des

45. Bien connu dans la littérature, le cosinus entre deux vecteurs numériques $\overrightarrow{v}$ et $\overrightarrow{w}$ se calcule comme suit (Jurafsky et Martin, 2009, 699) :

$$sim_{cosinus} = (\overrightarrow{v}, \overrightarrow{w}) = \frac{\overrightarrow{v} \cdot \overrightarrow{w}}{\mid \overrightarrow{v} \mid\mid \overrightarrow{w} \mid} = \frac{\sum_{i=1}^{N} v_i * w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}} \tag{3.19}$$

46. Cette recherche a d'ailleurs fait l'objet de deux articles publiés par ailleurs (François et Watrin, 2011a,b).

résultats satisfaisants (Evert et Krenn, 2001; Pearce, 2002), elles sont conçues pour extraire des UPs au départ d'un large corpus et non de contextes réduits. Elles reposent sur des mesures d'association (MAs), telles que le rapport de la log-vraisemblance de Dunning (1993), qui nécessitent de disposer d'une masse critique de données pour être estimée correctement. Par ailleurs, il n'est pas non plus possible de travailler sur la base d'un dictionnaire d'UPs, tant est grande la variabilité lexicale et syntaxique de ces unités.

L'extracteur : Pour faire face à cette limitation des approches classiques, nous avons défini par ailleurs (Watrin et François, 2011) une stratégie innovante qui permet de contourner cette difficulté à l'aide d'une base fréquentielle de n-grammes, que nous désignons aussi sous le terme de référence. Celle-ci autorise un calcul efficace des MAs à la volée, même pour des contextes réduits, pour autant que les fréquences stockées dans la base aient été dénombrées sur un large corpus. Nous avons repris les deux bases de n-grammes décrites précédemment (Google et *Le Soir*) pour constituer deux références.

Une fois cette étape effectuée, nous avons pu définir la procédure d'extraction des UPs en elle-même. Celle-ci s'inspire largement des travaux de Daille (1995) et de Smadja (1993) en ce qu'elle associe un ensemble de filtres linguistiques à la validation statistique basée sur les MAs. La présence de ces filtres permet de contraindre la nature grammaticale des termes candidats, ce qui assure une plus grande précision. Par contre, leur conception requiert un certain temps de développement. C'est pourquoi nous nous sommes limités dans cette étude aux structures nominales [47].

Concrètement, les textes sont d'abord annotés en termes de catégories de discours à l'aide du TreeTagger, avant que l'ensemble des termes candidats ne soit extrait au moyen des filtres linguistiques. Ceux-ci prennent la forme de transducteurs à états finis qui sont appliqués au texte cible à l'aide du logiciel *Unitex* (Paumier, 2003). Ensuite, cette liste de candidats est soumise au module de validation statistique qui assigne une valeur d'association à chacun d'eux. Dans cette version du système, la validation statistique implémente le rapport de la log-vraisemblance tel que décrit dans Silva et Lopes (1999) :

$$\begin{aligned} LogLik_f(w_1 \cdots w_n) &= 2 * \log L(pf1, kf1, nf1) + \log L(pf2, kf2, nf2) \\ &\quad - \log L(pf, kf1, nf1) - \log L(pf, kf2, nf2) \end{aligned} \tag{3.20}$$

Cette version du rapport de la log-vraisemblance présente l'avantage de considérer les UPs

47. Nos filtres considèrent les 4 structures nominales basiques suivantes : adjectif-nom (AN), nom-adjectif (NA), nom-nom (NN) et nom-préposition-nom ($NPrepN$). Par ailleurs, nous prenons également en compte la possibilité que trois transformations viennent s'appliquer à ces structures de base : la modification (typiquement sous la forme d'une insertion adjectivale ou adverbiale), la coordination et la juxtaposition (ex. $NPrepNPrepN$).

de taille supérieure au bigramme, tout en conservant de bonnes performances, ainsi que nous l'avons montré (Watrin et François, 2011).

Reste toutefois un dernier problème à envisager : quel critère utiliser pour décider qu'un terme candidat est effectivement une UPs. Le ratio de la log-vraisemblance de Dunning (1993) étant distribué selon une loi chi-carré à un degré de liberté, une première approche possible consiste à effectuer un test de significativité au départ de la valeur obtenue. Pour peu que celle-ci dépasse $3,84$, pour un $\alpha = 0,05$, on pourrait retenir le candidat. Malheureusement, à mesure que la taille du corpus de référence augmente, cette solution perd de son sens, car les hautes fréquences des occurrences engendrent des scores élevés pour le chi-carré. De plus en plus de phénomènes apparaissent alors significatifs (Kilgarriff, 2005). Nous montrons ici comme exemple les valeurs obtenues pour les trois mêmes collocations au départ de quatre corpus de taille croissante : trois collections d'articles du *Soir* (respectivement 500K, 1000K et 5000K mots) et le corpus de Google.

MWE	500 K	1000 K	5000 K	Google
même groupe	0,73	1.44	3,85	1 746,03
nouveaux instruments	3,81	3,3	49.83	2 793,65
lettres de noblesse	33,99	52,43	232,51	27 202,17

Tableau 3.1 – Exemples de deux candidats dont le ratio de la log-vraisemblance n'est pas significatif sur un petit corpus et qui le devient à l'extrême sur un très large corpus. Ils sont comparés à un troisième candidat qui est une véritable collocation.

La solution qui est communément apportée à ce problème est de fixer un seuil, non plus à partir des tables de la distribution chi-carré, mais empiriquement. Elle présente un défaut évident : ce seuil n'est valable que pour un corpus donné ou, à la rigueur, pour un corpus de taille comparable. Une fois de plus, l'usage d'une référence, dont la taille est constante, contourne cette difficulté et permet de fixer une fois pour toute un seuil optimal (dans notre cas, le choix du seuil a été opéré en fonction de la précision de l'extracteur) [48]. La valeur des seuils retenus est précisée dans le Chapitre 5.

La définition des variables : À partir de cet extracteur, il a été possible de définir 22 variables à même de prendre en compte diverses facettes des UPs :

NCPW : la proportion d'UPs nominales par rapport au nombre de mots.

48. Nous avons également expérimenté une troisième technique : le LocalMax (Silva *et al.*, 1999). Elle permet de se passer de seuil et retient parmi les candidats les collocations qui constituent un maximum local, c'est-à-dire dont le score d'association est supérieur à tous ceux de ces voisins de taille $n-1$ et $n+1$. Le LocalMax n'a toutefois pas apporté d'amélioration significative par rapport à un simple seuil, tandis qu'il augmente considérablement la charge de calcul. Nous ne l'avons donc pas retenu pour cette étude.

MeanSize : la taille moyenne (en nombre de mots) des UPs nominales présentes dans le texte.

NNColl : la proportion d'UPs nominales dont la structure grammaticale correspond à la suite de catégories *N N* ;

NPNColl : la proportion d'UPs nominales dont la structure est *N PREP (DET) N* ;

ANColl : la proportion d'UPs nominales constituées d'un adjectif suivi d'un nom (*A N*) ;

NAColl : la proportion d'UPs nominales nominales de structure *N A* ;

MPColl : la probabilité moyenne des UPs nominales rencontrées dans le texte ; les probabilités étant estimées sur la base des deux références.

MedPColl : la médiane des probabilités des UPs nominales rencontrées dans le texte ;

PColl75 : la probabilité correspondant au 75e percentile de la distribution des probabilités des UPs nominales du texte [49] ;

PColl90 : la probabilité correspondant au 90e percentile de la distribution des probabilités des UPs nominales du texte ;

MP2Coll : la probabilité moyenne des UPs nominales de taille 2 ;

MedP2Coll : la médianne des probabilités des UPs nominales de taille 2 ;

P2Coll75 : la probabilité correspondant au 75e percentile de la distribution des probabilités des UPs nominales de taille 2 ;

P2Coll90 : la probabilité correspondant au 90e percentile de la distribution des probabilités des UPs nominales de taille 2 ;

MP3Coll : la probabilité moyenne des UPs nominales de taille 3 ;

MedP3Coll : la médianne des probabilités des UPs nominales de taille 3 ;

P3Coll75 : la probabilité correspondant au 75e percentile de la distribution des probabilités des UPs nominales de taille 3 ;

P3Coll90 : la probabilité correspondant au 90e percentile de la distribution des probabilités des UPs nominales de taille 3 ;

MP4Coll : la probabilité moyenne des UPs nominales de taille 4 [50] ;

MedP4Coll : la médianne des probabilités des UPs nominales de taille 4 ;

P4Coll75 : la probabilité correspondant au 75e percentile de la distribution des probabilités des UPs nominales de taille 4 ;

P4Coll90 : la probabilité correspondant au 90e percentile de la distribution des probabilités des UPs nominales de taille 4.

Soulignons que ces facteurs ont pu être calculés à la fois sur avec Google et *Le Soir* comme référence, ainsi que différents seuils, ce qui donne un nombre important de variables, que détaillera davanatage le Chapitre 5.2.7.

49. Étant donné que les UPs les moins fréquentes se voient octroyées des probabilités plus petites, les termes complexes sont en réalité situés au début de la distribution. Toutefois, afin de conserver une cohérence globale pour nos diverses variables basées sur les percentiles, nous voulons que le terme « 75e percentile » désigne toujours les éléments les plus complexes de la distribution. Lorsque c'est nécessaire, comme dans ce cas-ci, il correspond en réalité au 25e percentile.

50. Nous avons limité à 4 tokens la taille des UPs prises en compte. Cela s'explique à la fois par la trop faible présence d'unités de taille supérieure dans les textes, ainsi que par la limitation imposée par les 5-grammes de Google.

Type de texte

Pour déterminer si le texte à évaluer est un dialogue ou pas, voire s'il contient une partie dialoguée, nous nous sommes inspiré des travaux d'Henry (1975) qui a mis en avant une série d'éléments linguistiques susceptibles de permettre la détection de ce type de textes, supposés plus faciles à lire. Sur cette base, nous avons défini les 5 variables suivantes :

PPD : le pourcentage de pronoms personnels de dialogue (c'est-à-dire de la 1re et 2e personne, à savoir : je, me, moi, tu, te, toi, nous, vous, ainsi que les pronoms possessifs : (le) mien, (le) tien, (le) nôtre, (le) vôtre) ;

PI : le pourcentage d'interjections dans le texte ;

PPEI_1 : le pourcentage de points d'exclamation et de points d'interrogation par rapport au nombre total de signes de ponctuation compris dans la liste suivante : (! ? . ;) ;

PPEI_2 : le pourcentage de points d'exclamation et de points d'interrogation par rapport au nombre total de signes de ponctuation compris dans la liste suivante : (! ? .) ; :) [51] ;

BINGUI : la présence de guillemets de dialogue (il s'agit d'une variable binaire).

Les informations nécessaires pour calculer ces variables ont été obtenues, une fois de plus, sur la base du TreeTagger, qui est capable de détecter les fins de phrases plus ou moins correctement, et signale les interjections (balise *INT*) et les guillemets (balise *PUN :cit*).

3.3.5 Conclusion

Avant de refermer cette seconde partie du chapitre sur les indices de la lisibilité, avec laquelle débute la partie expérimentale de notre étude, rappelons que pas moins de 406 variables y ont été définies. Celles-ci se répartissent dans quatre catégories, dont les plus peuplées sont les niveaux lexicaux (avec 170 facteurs) et syntaxiques (191), tandis que la dimension sémantique est celle qui a été la plus ardue à explorer, par faute de modèles théoriques, mais surtout par manque de ressources lexicales et humaines. Nous avons cependant réussi à implémenter 18 facteurs couvrant trois dimensions sémantiques identifiées dans la première partie théorique de ce chapitre : le niveau d'abstraction des textes, la difficulté conceptuelle des contenus et la cohérence des idées. Enfin, nous avons pris en compte 27 aspects spécifiques au processus de lecture en L2, en particulier la présence d'UPs.

Parmi ces nombreuses variables, la majorité a déjà fait l'objet d'études précédemment. Nous les avons répliquées afin de tester leur pertinence dans le contexte spécifique du FLE, un cadre dans lequel bien peu d'entre elles ont fait l'objet d'expérimentations. Ce seul point nous paraît déjà justifier l'intérêt de cette thèse qui propose probablement l'inventaire le plus complet des variables utilisables pour prédire la lisibilité de textes de FLE. Le fait d'avoir

51. Cette alternative à **PPEI_1** repose, d'une certaine manière, sur la notion de sous-phrase. Elle vise à défavoriser les textes comprenant de longues phrases, puisque, comme il l'a été dit plus haut, les dialogues ont tendance à présenter des phrases plus simples en moyenne.

répliqué autant de facteurs issus d'un grand nombre d'études va permettre de poser un discours critique sur certaines questions transversales en lisibilité que nous avons introduites dans le Chapitre 2.

Toutefois, il n'est pas inutile d'appuyer le fait que certaines des variables implémentées n'ont, à notre connaissance, tout simplement jamais été employées dans le contexte de la lisibilité. C'est le cas du nombre de voisins orthographiques, du TTR normalisé, de la plupart des variables basées sur les UPs, de la densité des idées mesurées à l'aide de *Densidées*, du nombre d'absents du vocabulaire d'*Alter Ego* 1 et 2 et des temps verbaux. Nous espérons évidemment découvrir parmi celles-ci l'un ou l'autre bon prédicteur, capable d'apporter une amélioration significative des performances après que les indices classiques ont été introduits dans le modèle.

3.4 Conclusion générale

Dans ce chapitre consacré à l'identification d'indices de lisibilité qui soient à la fois efficaces et implémentables, nous avons tout d'abord longuement détaillé ce qui s'est fait jusqu'à aujourd'hui, aussi bien pour l'anglais que pour le français, en L1 comme en L2. Cet appareil prédictif a été organisé en cinq classes : les variables lexicales, les syntaxiques, celles qui dénotent la difficulté conceptuelle, celles qui mesurent la cohérence textuelle et, enfin, celles qui sont spécifiques au contexte L2. Ce travail de catégorisation s'inspire de synthèses antérieures (Chall, 1958; Gray et Leary, 1935; Klare, 1963), tout en présentant des particularités. Parmi celles-ci, relevons la séparation qui a été faite entre deux aspects sémantiques (le contenu et son organisation), mais surtout la présence de sous-catégories de facteurs n'ayant jamais été discutées dans le cadre de la lisibilité.

Cependant, ce n'est pas tant l'utilité de cette classification ou l'apport de nouveaux prédicteurs qui nous semblent constituer la contribution la plus intéressante de ce chapitre, mais bien le soin minutieux qui a été apporté à discuter la pertinence de chacune des familles de variables à l'aide de données empiriques issues de la psycholinguistique et des neurosciences. Même si l'état actuel des connaissances ne permet pas encore d'être catégorique pour l'ensemble des facteurs de la difficulté traités ici, cette exploration dans une discipline connexe a révélé que, contrairement à ce que l'on a longtemps pensé en lisibilité, certains indices tels que les fréquences lexicales ou la longueur des mots apparaissent bien comme des déterminants influençant le processus de lecture. Rien n'empêcherait donc, sur base d'un tel inventaire, de revivifier le domaine si décrié qu'est la « writability ».

Cette analyse de la littérature a également mis en évidence les différents apports des technologies TAL à la lisibilité, ce qui constitue une question centrale de notre réflexion. Ceux-ci

ne sont cependant pas aussi évidents qu'ils pourraient le paraître au premier abord. En effet, nous avons vu que certaines variables, bien que reposant sur de solides bases théoriques, pouvaient se révéler de médiocres prédicteurs à la suite d'approximations ou d'erreurs en cascades dans les différents modules TAL mobilisés pour paramétriser des variables complexes. Parmi celles-ci, on retrouve en particulier les facteurs structuro-cognitivistes, dont l'intérêt pratique est plus discutable que ce que les discours théoriques auraient pu laisser croire. Plusieurs d'entre eux, que nous avons décrits, obtiennent en effet des coefficients de corrélations assez faibles ou, du moins inférieurs à ceux de prédicteurs classiques. Nous allons donc tenter, à l'aide de nos propres données, d'apporter un éclairage supplémentaire à ces questions dans la troisième et dernière partie de cette thèse, laquelle est consacrée à l'analyse des performances que les variables que nous avons décrites dans la seconde partie de ce chapitre obtiennent sur notre corpus de textes issus de manuels de FLE.

CHAPITRE 4

CORPUS D'ENTRAÎNEMENT : COMMENT MESURER LA DIFFICULTÉ DES TEXTES ?

4.1 Introduction

> *The problem of validity is difficult. First of all, there are no rigorous standards of just what is 4th grade difficulty as opposed to 5th grade difficulty. There seems to be some loose sort of agreement between publishers and educators which is based on experience and perhaps a little on test data as to what grade level designations mean. However, even standardized test data are not exact. Anybody who has used an old reading test, say the 1957 California Reading Test, on his class, then used the 1965 Stanford Reading Test on exactly the same class at nearly the same time, can tell you that the class mean reading score expressed in grade level is quite different. In general, newer test are more difficult or, in other words, a 9th grade student today reads better than a 9th grade student in former years. Fry (1968, 515)*

La conception d'une formule de lisibilité ne se limite pas à implémenter un ensemble d'indices de la difficulté sélectionnés avec soin. Il faut également rassembler un ensemble de textes dont la difficulté est connue et exprimée selon une échelle de référence. Ces textes pourront ensuite être mesurés sous toutes leurs coutures afin d'obtenir une valeur pour chacun des prédicteurs envisagés. Ils servent ainsi de points de référence pour prédire le niveau de nouveaux textes. Par exemple, si l'on observe que les textes annotés comme faciles comportent en moyenne des phrases plus courtes que les autres textes, on peut en déduire que de

nouveaux textes comportant des phrases courtes ont une certaine probabilité d'être simples aussi.

Ces régularités, que l'on peut exprimer comme des règles de décision, peuvent être obtenues au moyen d'une analyse manuelle, plutôt fastidieuse à réaliser. C'est pourquoi elles sont souvent inférées automatiquement à partir d'un corpus de textes dont on connaît à la fois la difficulté et les caractéristiques. On appelle **classification** ce type d'approche (Sebastiani, 2002; Tan *et al.*, 2005, cf.), communément utilisée en TAL et que la lisibilité computationnelle s'est mise à employer (Heilman *et al.*, 2008; Schwarm et Ostendorf, 2005; Si et Callan, 2001). Toutefois, si l'emploi d'un classifieur évite au chercheur d'avoir à définir des règles de décision, le retour de bâton ne se fait pas attendre. En effet, pour entraîner ce type de modèles, il est nécessaire de disposer d'un large corpus où la variable à prédire (ici, la difficulté) est annotée. Dans notre cas, celui-ci permettra d'apprendre automatiquement les relations qui existent entre la difficulté des textes et les caractéristiques linguistiques décrites dans le Chapitre 3.

Ce chapitre décrit comment nous avons constitué un tel corpus annoté, ce qui nous a demandé de

1. sélectionner un **critère** – le plus objectif possible – permettant d'évaluer la difficulté de textes pour la population ciblée ;

2. rassembler un ensemble de textes, annotés en fonction du critère retenu, et dont les caractéristiques (date de rédaction, genre, sujets traités, etc.) sont représentatives des textes que l'on désire évaluer ensuite ;

3. opter pour une échelle de référence, aussi pratique que possible, et qui puisse être reliée au critère.

Étant donné que la sélection du critère détermine la manière dont les textes seront collectés et oriente le choix d'une échelle de difficulté, nous ouvrons ce chapitre en discutant les différents critères possibles et le choix que nous avons opéré (*cf.* Section 4.2). La seconde section (*cf.* Section 4.3) décrit le corpus en lui-même : les critères de sélection des textes, leurs caractéristiques et le format employé pour annoter le corpus. Enfin, la dernière partie (*cf.* Section 4.4) rapporte les résultats de diverses analyses visant à évaluer la pertinence du corpus rassemblé pour concevoir une formule de lisibilité.

4.2 Le choix du critère

4.2.1 Introduction : qu'est-ce qu'un critère ?

En de nombreux endroits de cette thèse, le lecteur a pu rencontrer des expressions telles que « mesurer la difficulté des textes », « textes annotés en fonction de leur difficulté », etc. Nous n'avons pas estimé nécessaire de définir le terme « difficulté », puisque cette notion est aisément comprise par un humain, comme nous l'avons souligné au début du Chapitre 2. Toutefois, lorsqu'il devient nécessaire de la circonscrire précisément afin de la traduire sous une forme numérique, on se rend compte qu'il s'agit d'une notion fuyante, qui échappe à la description et recouvre une réalité multiple.

Malheureusement, l'approche pragmatique qui dirige les études en lisibilité ne peut rien tirer de cette multiplicité floue. Il lui faut dès lors l'aborder de biais, en identifiant une autre variable, plus facile à mesurer et fortement corrélée avec ce que l'on considère comme la difficulté. C'est cette autre variable qu'on appellera ici un **critère**. Dans la suite de cette section, nous allons décrire les critères qui ont été envisagés jusqu'ici dans les études de lisibilité, à savoir l'avis d'experts, les tests de compréhension, le test de closure, la vitesse de lecture, le rappel, ainsi qu'une série de critères plus marginaux.

4.2.2 L'avis d'experts

Le recours à des experts (professeurs de langue, concepteurs de manuels et autres pédagogues) a très certainement été la première méthode employée pour classer des textes en fonction de leur difficulté. Des juges évaluent la difficulté des textes cibles en se projetant dans la peau de lecteurs potentiels dont ils connaissent les caractéristiques à force de les fréquenter ou de les éduquer. Pour Henry (1975, 50), cet aspect projectif ne doit pas être oublié et l'adéquation entre la difficulté réellement éprouvée par les lecteurs et les jugements fondés sur des critères d'adultes ou de natifs n'est pas évidente. Il prend pour preuve une étude de de Landsheere (1978) dans laquelle six textes de difficulté variable ont été étalonnés par vingt experts. Leurs avis sont sujets à une variation que Henry juge trop importante pour valider l'approche par experts.

Toutefois, cet avis critique est loin d'être partagé par tous les chercheurs en lisibilité. Plusieurs d'entre eux ont effectué des tests confirmant la pertinence de ce critère ou l'ont employé pour développer une formule. Nous avons déjà détaillé les expériences de Carver (1974a, 1975) et de Singer (1975) (*cf.* Chapitre 2.2.5). Ces auteurs démontrent d'une part qu'un nombre limité de juges peuvent s'accorder relativement bien sur l'annotation d'un ensemble de textes pour peu qu'ils aient préalablement été reconnus comme experts et, d'autre

part, que les prédictions réalisées par ce groupe d'experts sont valides.

Similairement, Björnsson (1983, 482) obtient une corrélation de Spearman de $\rho = 0,97$ entre les avis de deux groupes d'experts sur un même corpus. Il conclut :

> *Traditionally it has been thought that judges' ratings of absolute difficulty are unreliable. From our experience they are not, i.e., when they are made by a fairly large group of persons, when the passages are relatively long, and when the range in difficulty in the text battery is wide.*

Par ailleurs, Henry (1975, 46) répertorie pas moins de 15 études ayant utilisé l'avis d'experts pour valider des formules de lisibilité et la plupart d'entre elles présentent un coefficient de corrélation significatif entre les deux méthodes (formule et jugement humain).

À côté de ces expériences visant à évaluer directement la qualité des jugements humains, il faut souligner que ce critère a été utilisé par de nombreux concepteurs de formules, depuis Lively et Pressey (1923). Citons entre autres Dolch (1948), Wheeler et Smith (1954), Spache (1953), dont la formule a été reconnue pour sa grande validité, ou encore Elley (1969). Plus récemment, il est devenu le critère phare du courant de la lisibilité computationnelle et se retrouve dans la majorité des études actuelles (Collins-Thompson et Callan, 2004a; Pitler et Nenkova, 2008; Schwarm et Ostendorf, 2005; Si et Callan, 2001).

Cet engouement n'est pas étonnant. En effet, recourir à des experts mobilise normalement un certain nombre de ressources temporelles et humaines. Cependant, comme ce type de critère est aussi utilisé pour la conception de manuels scolaires ou de pages web à vocation pédagogique, il est dès lors aisé de rassembler un grand nombre de textes à partir de ces sources déjà annotées, ce qui rencontre parfaitement les besoins des approches TAL actuelles (*cf.* Section 4.2.8).

Toutefois, comme nous l'avons expliqué dans la conclusion du chapitre 2, la pertinence de ce critère est de moins en moins questionnée, à mesure que la lisibilité passe des mains du monde de l'éducation à celles du monde de l'ingénierie, où les questions de psychométrie sont probablement moins prégnantes. Soulignons toutefois l'étude de van Oosten *et al.* (2011), qui pose la même question qu'Henry à l'aide de techniques statistiques modernes. Les auteurs font évaluer, par paires, 105 fragments textuels par une série d'experts. Ils regroupent ces derniers selon la similarité de leurs annotations via un algorithme de clustering. À chaque groupe d'experts correspond un ensemble de textes, subdivisés en un corpus d'entraînement et un de test. Ensuite, pour chaque corpus d'entraînement ainsi disponible, ils entraînent des modèles de classification binaire dont les performances sont évaluées à la fois sur le corpus de test de leur cluster (validation intra-cluster) et sur les corpus de tests issus des autres groupes d'experts (validation inter-cluter). Un biais apparaît dès lors : les

performances des différents modèles diminuent sensiblement en validation inter-cluster, ce qui conduit les auteurs à douter de la possibilité d'atteindre un accord interannotateurs satisfaisant.

Ces divergences dans les avis d'experts s'expliquent notamment par l'obligation qu'ont les juges de se mettre à la place des lecteurs cibles. Ils réalisent plus ou moins bien cette opération en fonction de leur expérience, mais peuvent également diverger par la façon dont ils considèrent le fait de comprendre un texte. Autrement dit, la question essentielle qui se pose lorsqu'on fait appel aux jugements d'experts est de déterminer les critères qui serviront à les qualifier comme tels. Carver (1975) et Singer (1975) ont montré l'intérêt d'une étape de sélection. De leur côté, van Oosten et Hoste (2011) ont montré qu'il était valide de remplacer les experts par des lecteurs issus de la population à modéliser, pour autant qu'on en dispose d'un nombre suffisant, notamment en employant des techniques de *crowdsourcing*.

4.2.3 Tests de compréhension

Confrontés aux limites des évaluations réalisées par des spécialistes, les premiers chercheurs en lisibilité (Dale et Tyler, 1934; Ojemann, 1934) ont envisagé une seconde approche : tester la compréhension directement auprès des lecteurs, afin d'éviter le prisme déformant de l'expert. Le niveau de difficulté d'un texte correspond alors à la moyenne des résultats obtenus par les sujets testés. Cette approche présente l'énorme avantage de mesurer directement l'interaction entre le texte et le lecteur, propre au processus de lecture (*cf.* Chapitre 1). La situation de lecture, troisième dimension de ce processus, est quant à elle contrôlée : tous les lecteurs disposent du même temps, reçoivent les mêmes consignes et, dans la mesure du possible, sont rassemblés en même temps dans le même lieu.

Apparaissant a priori comme le meilleur critère, les tests de compréhension ont pourtant fait l'objet de nombreuses critiques, qu'ils soient constitués de questions ouvertes ou à choix multiples. Ils présentent diverses faiblesses qui remettent en question leur validité comme mesure de la compréhension d'un texte. La plus cruciale d'entre elles est sans doute l'interaction qui existe entre le texte à évaluer et les questions conçues à cet effet. Lorge (1949, cité par Chall, 1958, 40) souligne le fait que les questions sont porteuses d'une difficulté propre qui est amalgamée à celle des matériaux à tester. Il est alors possible qu'un sujet, confronté à un texte simple, soit dérouté par une question formulée en des termes nébuleux ou, à l'inverse, qu'une bonne réponse à une question simple cache un problème de compréhension par rapport au contenu. Davis (1950, cité par de Landsheere, 1978, 33) a confirmé cette hypothèse à l'aide de l'expérience suivante. Il a conçu deux versions d'un questionnaire à choix multiples, l'un rédigé avec des mots fréquents, l'autre avec des mots peu fréquents. Il les a soumis à deux groupes de sujets et a observé une différence élevée entre la réussite de

l'un et l'autre.

S'il n'est déjà pas aisé d'accorder la complexité d'un texte et d'une série de questions, que dire lorsqu'on veut comparer plusieurs textes ? Parvenir à produire des séries de questions, portant sur des textes différents, et dont la difficulté est équivalente relève de la gageure. C'est d'autant plus laborieux dans le cas de questionnaires à choix multiples, où s'ajoute à la difficulté linguistique des questions celle induite par le type et la quantité de distracteurs. Une troisième objection émise à l'encontre des tests de compréhension est leur incapacité à porter sur l'ensemble du texte, le nombre de questions étant généralement insuffisant pour ce faire (de Landsheere, 1978). Par ailleurs, Gilliland (1972, 88) fait remarquer que l'ordre dans lequel les questions sont présentées importe.

Au niveau de la correction, l'emploi de tests de compréhension génère également certains problèmes. Comment assurer l'objectivité de la méthode de dépouillement dans le cas de questions ouvertes ? Comment, surtout, interpréter les scores obtenus ? Chall (1958) explique qu'un résultat identique, par exemple 60% de bonnes réponses, ne signifie pas la même chose pour un enfant du début du primaire – dont ce sont surtout les capacités de décodage qui sont évaluées – et un adolescent terminant ses secondaires. Autrement dit, si ce critère permet de classer un ensemble de textes les uns par rapport aux autres, une étape supplémentaire est nécessaire pour faire le lien entre les scores obtenus à un test et un niveau d'éducation donné. Sans compter la dépense d'énergie pour concevoir l'ensemble des tests et le temps requis pour les soumettre à un échantillon de taille suffisante.

Malgré ces nombreuses difficultés méthodologiques, les tests de compréhension ont dominé la lisibilité durant les années 30 à 60, puisqu'il s'agit de la technique ayant servi à étalonner les leçons de McCalls and Crabbs, tant usitées durant la période classique (Dale et Chall, 1948a; Flesch, 1948; Gunning, 1952). Ces dernières présentaient un double avantage : non seulement les textes qu'elles comprenaient étaient annotés en termes de difficulté, mais le lien avec les *grades* américains avait déjà été fait (*cf.* Chapitre 2.2.3). Comme nous l'avons déjà dit aussi, ce critère a été battu en brèche après 1965, au profit du test de closure. Aujourd'hui, il n'est plus vraiment considéré comme valable pour la lisibilité (Benoît, 1986b, 57), bien qu'il soit encore parfois employé, par exemple par Feng *et al.* (2009).

4.2.4 Le test de closure

Pour Taylor (1953), il suffit de supprimer dans un texte un mot sur cinq et de demander à un sujet de combler ces lacunes pour mesurer le niveau de compréhension auquel il est arrivé pour ce texte. Cette technique, le test de closure, a été décrite en détail dans le Chapitre 2.2.5. Comme les tests de compréhension, elle comporte trois variables : la forme du

test, le lecteur et le texte dont on veut mesurer la difficulté [1]. Cette opération requiert de contrôler les deux premières variables, ce qui s'est révélé extrêmement ardu à réaliser dans les tests de compréhension classique. Or, la forme simple et systématique du test de closure rend ce contrôle plus aisé. Il est possible, pour deux personnes différentes, de produire exactement le même test et de le corriger de la même façon. Encore faut-il que le test de closure mesure bien la compréhension à la lecture ! Cette propriété, qui a fait l'objet de nombreux débats, est discuté dans la première partie de cette sous-section. Dans un deuxième temps, nous présentons les diverses critiques ayant été adressées à l'utilisation du test de closure en lisibilité.

Le test de closure comme mesure de la compréhension

Cette question a donc engendré de nombreux débats, qui s'expliquent en partie par la nature composite du test. Celui-ci mesure en effet plusieurs choses : ce que le lecteur connaît, la lisibilité d'un texte et sa redondance, c'est-à-dire son niveau de prédictibilité (Klare *et al.*, 1972). Dès lors, d'après Henry (1975, 54), il n'est possible d'établir la validité opérationnelle de ce test que via des méthodes corrélationnelles. Autrement dit, on valide le test de closure comme outil de mesure de la compréhension en montrant que ses résultats sont corrélés avec ceux de tests de compréhension classiques. Taylor (1957) effectue une série d'expériences dans ce but. À partir d'articles traitant de sujets militaires, il compare les résultats de tests de closure à ceux de questionnaires à choix multiples et obtient des corrélations situées entre $0,51$ et $0,92$. Par ailleurs, il montre que ces mêmes résultats sont également corrélés significativement avec les performances des sujets à un test d'aptitudes mentales. Dans la même veine, Jenkinson (1957, citée par Jongsma, 1969, 13) est l'une des premières à comparer le test de closure avec des tests de lecture standardisés. Elle note que, pour son public composé d'étudiants du secondaire, les scores de closure sont corrélés à $0,78$ avec la partie portant sur les connaissances lexicales des tests standardisés et à $0,73$ avec la section ciblant la compréhension. Cette association entre les deux types de tests est reproduite par Bormuth (1969b, 365) entre autres [2], qui estime ceux-ci équivalents.

Il existe cependant quelques études qui mettent en doute cette équivalence. Weaver et Kingston (1963) obtiennent ainsi des corrélations relativement faibles (de $0,25$ à $0,51$) entre

1. En se basant sur ce modèle à trois variables, Richaudeau (1984) distingue plus généralement trois applications possibles du test de closure en éducation :
   – Mesurer la performance de lecteurs, en partant d'un matériel linguistique préalablement étalonné ;
   – Mesurer le degré de lisibilité d'un texte, sur la base d'un groupe de lecteurs dont le niveau est connu ;
   – Concevoir ou valider des formules de lisibilité, à partir de textes et de sujets étalonnés.

.

2. Pour des raisons de place et de pertinence par rapport au sujet traité, nous ne présentons ici que quelques travaux en faveur du test de closure. Il en existe toutefois d'autres, cités par Jongsma (1969, 9-13), (de Landsheere, 1978, 38-41) ou Henry (1975, 55).

leurs tests de closure et le *Davis Reading Test*. Par ailleurs, Marlier (cité par Henry, 1975, 55) n'obtient qu'une corrélation de $0,44$ avec les tests de lecture I.E.A. Notons qu'il s'agit, dans ce dernier cas, d'un mémoire de licence. Quant à la première étude, ses auteurs ont construit des tests où les lacunes était constituées tantôt uniquement de termes lexicaux, tantôt uniquement de mots grammaticaux. Or, Taylor (1957) a montré que le mélange des deux engendre une meilleure corrélation, ce qui pourrait expliquer des associations plus faibles dans le cas de Weaver et Kingston (1963).

En résumé, il semble que la majorité des études se prononcent en faveur de l'équivalence entre les tests de compréhension et de closure, ce qui a conduit à un emploi fréquent de ce critère en lisibilité. Il présente en effet plusieurs avantages par rapport aux tests de compréhension (Klare *et al.*, 1972, 82–83) : (1) la facilité de construction, même lorsque le concepteur ignore tout du domaine traité ; (2) l'absence de problèmes d'équivalence entre la difficulté du texte et des questions ; (3) la simplicité de correction ; (4) la possibilité de construire aisément des versions équivalentes à partir d'un même texte et (5) la prise en compte des caractéristiques du lecteur, telles que sa maturité intellectuelle, ses connaissances préalables, sa familiarité avec le sujet et son bagage culturel.

Critiques

Le fait qu'une majorité d'études attestent de la validité du test de closure comme mesure de la compréhension n'a pas empêché ses détracteurs d'en souligner certaines limitations, qui sont parfois vues comme des forces par d'autres. Ainsi, un premier aspect polémique est l'incapacité du test de closure à cibler une compétence en particulier – ce que permet un test de compréhension. Henry (1975, 58) considère toutefois cette caractéristique comme un atout, estimant que le test de closure constitue une tâche plus globale qui demande à la fois de « tirer parti d'informations lexicales ; être capable d'appréhender certaines structures syntaxiques ; se dégager du contexte immédiat et repérer les références lointaines. ».

Cette dernière compétence est toutefois loin d'être avérée. En réalité, un second reproche régulièrement adressé au test de closure est justement le fait qu'il ne nécessite que rarement de dépasser le contexte phrastique pour répondre, contrairement à ce que prétend Henry. Utilisant un protocole où des sujets doivent deviner, les uns après les autres, chacun des 150 mots de fragments textuels, Miller et Coleman (1967) montrent que les réponses produites lors de cette procédure sont peu contraintes par les phrases qui précèdent celle où se situe le mot à deviner. Ce serait donc les informations intraphrastiques qui importeraient le plus à la lecture. Cette observation est confirmée par Bartoo (1976, cité par Shanahan *et al.*, 1982, 233) qui, dans sa thèse, observe que l'ajout d'un contexte non lacunaire avant le fragment lacunaire n'améliore pas les performances de sujets sur ce dernier. Quant à Carver (1974b),

il envisage un autre traitement de la question, mais parvient à des conclusions similaires. Comparant les scores de closure de sujets ayant lu préalablement le texte utilisé comme support avec ceux de sujets ne disposant pas de cet avantage, il note que la différence est minime. Cela semble indiquer que les résultats à des tests de closure sont peu influencés par les informations obtenues lors de la lecture [3].

À l'inverse, certaines études défendent l'idée qu'il est nécessaire d'utiliser des informations interphrastiques pour répondre correctement. Étienne (1971, citée par Henry, 1975, 59) note que la majorité de ses sujets ne cherchent pas à comprendre un paragraphe dans sa globalité avant de compléter les lacunes qu'il contient. Pourtant, ceux qui effectuent cette opération obtiennent les meilleurs résultats. L'année suivante, Ramanauskas (1972) effectue une expérience où il compare les performances entre une version normale d'un texte (dont il propose une version lacunaire) et une version lacunaire de ce même texte où les phrases ont été mélangées, afin de briser les relations entre phrases. Comme il l'espérait, il observe de meilleurs scores pour les textes normaux.

Confrontés à ces diverses études contradictoires, Shanahan *et al.* (1982) ont effectué trois expériences supplémentaires dans le but d'essayer de trancher la question. La première et la seconde répliquent le mode opératoire de Ramanauskas (1972), mais contrairement à ces derniers, ne font apparaître aucune différence significative entre les résultats sur les textes normaux et sur ceux où les phrases ont été mélangées, même lorsque le temps accordé pour répondre est limité. La troisième expérience consiste à insérer une unique phrase tirée de 8 fragments lacunaires dans un environnement sémantiquement différent et également lacunaire. De la sorte, il est possible de comparer les performances sur cette même phrase lorsqu'elle est insérée dans son contexte ou perdue au milieu de phrases sans lien sémantique. Une fois de plus, cette expérience ne conduit pas à une différence significative entre les deux traitements, ce qui pousse les auteurs à rejeter l'hypothèse selon laquelle le test de closure serait influencé par l'utilisation d'informations interphrastiques. Il s'agit là évidemment d'une sérieuse faiblesse pour ce critère, en particulier lorsqu'il est appliqué à des lecteurs expérimentés.

Enfin, la dernière critique adressée au test de closure est sa tendance à plutôt mesurer la redondance du texte que le niveau de compréhension d'un lecteur (Kemper, 1983; Kintsch et Vipond, 1979; Richaudeau, 1973, 1981; Weaver et Kingston, 1963). En effet, Taylor (1955, cité par Coleman et Miller, 1968, 371) montre que les résultats de tests de closure sont inversement proportionnels au nombre de bits d'incertitude dans le passage, tels que mesurés dans le cadre de la théorie de l'information (Shannon et Weaver, 1949). Richaudeau est sans

3. Pour une discussion plus approfondie de la question et des liens vers d'autres études opposées à l'emploi du contexte interphrastique dans les tests de closure, se reporter à Shanahan *et al.* (1982).

doute l'un des plus fervents détracteurs du test de closure appliqué en lisibilité lorsqu'il affirme qu'alors « ces formules risquent de confondre par trop les notions de redondance et de lisibilité. » (Richaudeau, 1978, 109). Un test porteur d'une entropie nulle serait en effet le plus lisible des textes, mais quel serait son intérêt ? Il montre également que ce sont les mots outils, c'est-à-dire les catégories de discours les plus conditionnées par le contexte, qui sont les mieux reconstitués par les sujets. Toutefois, Miller et Coleman (1967, 854) défendent l'opinion que le test de closure est plus performant qu'une simple mesure de redondance telle que définie par Newman and Gerstman (1952, cités par Miller et Coleman, 1967, 854).

Dès lors, il nous semble qu'il faut nuancer cette troisième critique. Il est vrai que certaines lacunes des tests de closure sont très fortement conditionnées par le contexte direct et sont redondantes en ce sens qu'il est possible de comprendre le texte en leur absence (ex. : *Il était ..... fois un petit garçon .... avait de grandes mains.*). Mais, les lacunes mesurent aussi d'autres aspects : la connaissance des structures de la langue (ex. *un homme à ..... faire*), les connaissances culturelles ou conceptuelles (ex. *La ..... universelle des droits de l'..... a été signée en ..... à Paris.*), et aussi la compréhension du texte (ex. *il suffit de .... dans un texte un ..... sur cinq et de ..... à un sujet de ..... ces lacunes pour mesurer ..... compréhension*).

Pour terminer cette discussion sur le test de closure comme critère en lisibilité, il convient de souligner que son emploi dans le contexte des langues étrangères a encore été davantage sujet à caution que dans le cas de natifs. Oller et Conrad (1971) rapportent que dès 1968, Allen propose d'utiliser des textes à trous où les lacunes sont soigneusement choisies, tandis que Friedman défend dans sa thèse (publiée en 1974) l'intérêt du test de closure comme mesure de la compréhension à la lecture pour des apprenants d'ALS. C'est surtout suite aux études de Oller et Conrad (1971) et d'Oller (1972b), qui concluent que le test de closure mesure plutôt des compétences plus globales que simplement des aspects lexico-syntaxiques, que les professionnels des langues secondes et étrangères l'ont considéré comme un test universellement valide, aussi bien pour mesurer la compétence langagière que la compréhension (Alderson, 1979, 220).

Cet optimisme n'est pourtant pas partagé par tous (Alderson, 1979; Stevenson, 1981). Alderson (1979) effectue une expérience fort similaire à celle d'Oller (1972b), mais conclut à l'inverse : les résultats du test de closure seraient davantage associés aux parties lexico-syntaxiques d'un test de niveau (ici, l'ELBA) qu'aux parties visant à mesurer les capacités de compréhension à la lecture. Par ailleurs, l'auteur met aussi en évidence le fait qu'introduire des variations dans les tests de closure (le type de correction, la fréquence de suppression et la difficulté du texte support) conduit à des variations dérangeantes dans le type de compétences mesurées. Il termine ainsi son argumentation :

The procedure [here, the cloze test] is in fact merely a technique for producing

> *tests, like any other technique, for example the multiple-choice technique, and is not an automatically valid procedure. [...] Perhaps the principle of randomness needs to be abandoned in favour of the rational selection of deletions.*

Plus tranché dans son avis, Laroche (1979, 132) considère le test de closure comme un critère inadéquat pour des formules de lisibilité en L2. En effet, ce test est censé refléter l'acquisition d'une certaine intuition à propos du langage et Laroche estime qu'il n'y a pas de population de ce genre dans le cas d'une L2. Cette opinion est toutefois mise à mal par les résultats expérimentaux d'Oller et Conrad (1971) qui n'observent pas de différences de performances chez des apprenants avancés et des natifs, ce qui laisse penser qu'une telle intuition est bien acquise au bout d'un certain temps d'apprentissage.

Schulz (1981) avance un troisième grief : appliqué en L2, le test de closure inclut une dimension de production – en tout cas moins présente en L1 –, puisqu'il est possible qu'un apprenant devine la réponse requise dans sa langue, mais ne soit pas capable de retrouver l'équivalent dans la L2. Cette critique se voit atténuée dès lors qu'on n'accepte pas uniquement la forme correcte, mais aussi l'ensemble des mots acceptables en fonction du contexte. Oller (1972b) montre d'ailleurs que les résultats issus d'une telle correction sont davantage corrélés avec les performances au test *UCLA ESL Placement Examination*, dans le cas d'apprenants d'une langue étrangère [4].

Enfin, l'une des remarques les plus pertinentes est due à Greenfield (2004). Il souligne que l'on ignore à partir de quel score il est possible de considérer qu'un texte a été compris par le lecteur. Si, pour les natifs, Bormuth (1966) a effectué des expériences révélant qu'un score de 35% à 55% dénotait un taux de compréhension satisfaisant, aucune étude de ce type n'aurait été entreprise pour les L2. D'autant que les apprenants seraient plus tolérants à un décodage imparfait que les natifs, d'où la possibilité qu'ils comprennent correctement un texte donné sans pour autant être capables de bien réussir un test de closure sur ce même texte (Greenfield, 2004). Cet argument nous semble en partie discréditer l'usage de tests de closure pour la L2.

En conclusion, il apparaît que, malgré ses nombreux avantages pratiques, le test de closure n'a pas réussi à faire l'unanimité en ce qui concerne sa validité pour mesurer le niveau de compréhension d'un individu envers un texte donné. Par ailleurs, même s'il constitue une procédure de test simplifiée, il requiert néanmoins le recours à un large échantillon d'apprenants, en particulier lorsqu'il s'agit d'annoter une grande quantité de textes. Pour ces deux raisons, nous avons écarté le test de closure comme critère pour notre étude.

4. Par contre, pour des natifs, Miller et Coleman (1967) avaient démontré que cette forme de correction plus souple était corrélée à $0,99$ avec une correction où l'on n'accepte que la forme attendue, et qu'il ne servait à rien de l'appliquer.

4.2.5 La vitesse de lecture

D'après Klare (1963, 135), Flesch (1949) aurait été le premier à associer lisibilité et vitesse de lecture. Pourtant, ce serait oublier un peu vite l'étude de McClusky (1934) (*cf.* Chapitre 2.2.2) qui envisageait déjà de mesurer le temps mis par des sujets pour lire un texte avant de répondre à un questionnaire du type « Vrai ou Faux ». Le recours à un test en aval de la lecture permettait d'éviter que les sujets n'expédient la tâche, ce qui aurait biaisé les temps de lecture. Or, c'est justement cette donnée qui est finalement utilisée comme critère par McClusky (1934).

Ce critère alternatif n'a toutefois pas connu un grand succès par la suite. Il faut dire que tant que les techniques de chronométrage manquaient de précision, il apparaissait peu fiable. De plus, l'association entre la vitesse de lecture et la qualité de la compréhension n'a été établie que peu à peu. Flesch (1949) suggère effectivement que le nombre de mots lus par minute, mais aussi le nombre de fixations, de régressions et de clignements des yeux sont influencés par la difficulté des textes. Toutefois, il n'opère pas de lien avec la compréhension, laquelle n'est pas mesurée dans son étude.

Il faut attendre Brown (1952) pour que soit réalisée une expérience où la vitesse de lecture et le taux de compréhension sont mesurés et comparés. L'auteur utilise deux tests de lecture de difficultés différentes et les soumet à une centaine de sujets, dont la vitesse de lecture est aussi mesurée pour les textes utilisés dans ces tests. La vitesse moyenne de lecture sur un texte difficile s'élève à 306 mots/min. contre 235 mots/min. pour le texte très difficile. Le niveau de compréhension, mesuré à l'aide du test, laisse également apparaître une différence qui paraît sensible, même si son taux de significativité n'est pas spécifié. Ce résultat semble plaider pour l'emploi de la vitesse de lecture comme critère en lisibilité, à l'instar du niveau de compréhension.

Par la suite, cette association est confirmée empiriquement par Oller (1972a) et étayée théoriquement par Just et Carpenter (1980). Le premier observe des corrélations significatives entre plusieurs variables liées aux mouvements des yeux et les scores réalisés aux différentes sections d'un test de langue standardisé (il s'agit toujours de l'*UCLA ESLPE*). Cette fois, on ne note guère de différences entre les parties portant sur le vocabulaire, la syntaxe et la composition. Quant aux seconds, ils défendent l'hypothèse suivante : l'interprétation de ce qui est perçu à la lecture serait immédiate et, par conséquent, le temps de fixation oculaire d'un mot donné correspondrait en gros au temps de traitement cognitif de celui-ci (*cf.* Chapitre 1.2.3). Cette découverte semble également en faveur de l'emploi de la vitesse de lecture comme critère en lisibilité.

Malgré ces études favorables, la rapidité de lecture reste très peu employée en lisibilité.

D'une part, se limiter à mesurer la vitesse de lecture peut mener à des biais : sachant que leur rapidité de lecture est mesurée, les sujets pourraient être tentés de se contenter de lire mécaniquement, sans se préoccupper d'appréhender le contenu du texte. Dans ce cas, l'association entre temps de lecture et taux de compréhension n'est plus valide. C'est d'ailleurs, la raison qui empêche Henry (1975, 52) d'utiliser ce critère pour sa formule. Il distingue en effet la vitesse de compréhension de la simple vitesse de lecture, où l'on ne mesure que le mécanisme perceptif.

Plusieurs manières de se garantir contre ce risque de biais ont cependant été envisagées. La technique la plus répandue consiste à vérifier, d'une manière ou d'une autre, la compréhension après la lecture, soit à l'aide de tests (McClusky, 1934), soit via le taux de rappel (Kintsch *et al.*, 1975; Kintsch et Vipond, 1979; Miller et Kintsch, 1980). Par ailleurs, il est possible de combiner ces deux mesures pour définir un critère qui prenne à la fois en compte le temps de lecture et la compréhension. Kintsch *et al.* (1975) recommandent ainsi l'emploi du nombre moyen de secondes par proposition qui remémorée.

Une seconde raison qui explique sans doute le faible emploi de ce critère est son coût, tant au niveau matériel que temporel. En effet, la vitesse de lecture est généralement mesurée à l'aide des trois techniques d'oculométrie décrites au Chapitre 1.2.2, lesquelles requièrent l'emploi d'appareils onéreux. Par ailleurs, comme il est nécessaire de mesurer cette variable chez chaque sujet et pour chaque texte, on imagine aisément les efforts requis pour constituer un corpus d'une taille suffisante.

Face à ces problèmes pratiques, l'auto-présentation segmentée (APS) pourrait constituer une solution intéressante. Dans cette approche, le sujet lit un texte via un programme informatique qui permet la manipulation suivante : le texte à lire « est présenté à l'écran [...], de sorte que le lecteur ne voit à la fois qu'une partie plus ou moins restreinte du texte. À chaque fois que le lecteur appuie sur un bouton, le segment suivant lui est présenté » Reinwein (1992, 264). Il est alors possible de mesurer le temps pris pour chaque segment. Cette procédure présente cependant deux défauts. D'une part, la taille des segments influence la lecture : « Si des unités linguistiques plus larges favorisent un type de lecture plus proche d'une lecture dite normale, des unités linguistiques plus petites, par contre, permettent une analyse plus détaillée des temps de lecture, l'écologie de la tâche expérimentale et le raffinement des résultats étant diamétralement opposés » Reinwein (1996, 265). D'autre part, elle ne dispense pas le chercheur de vérifier le niveau de compréhension des sujets, afin d'éviter une lecture superficielle.

Reinwein (1992) suggère une variante intéressante de l'APS : la technique ZigZag. Celle-ci « oblige le lecteur a choisir continuellement entre deux mots dont un seul est parfaitement acceptable dans le contexte des segments textuels déjà présentés et à venir » Reinwein (1996,

148). Il s'agit donc d'une sorte de QCM permanent, qui permet de finement vérifier la bonne compréhension, mais présente le défaut de ralentir considérablement la lecture. La tâche demandée aux sujets s'éloigne donc sensiblement d'une lecture libre, se rapprochant du test de closure si ce n'est que le temps est ici aussi considéré. Bien que ce critère semble intéressant, nous ne l'avons pas employé, faute de disposer d'un tel système.

4.2.6 Le rappel

Un quatrième critère utilisable pour constituer un corpus de textes annotés est le taux de rétention ou de rappel, c'est-à-dire la quantité d'idées du texte dont un lecteur donné parvient à se rappeler, soit immédiatement après la lecture, soit à la suite d'un délai plus ou moins long. Il est difficile de déterminer qui a suggéré le premier ce critère en lisibilité. Dale et Tyler (1934, 386) l'envisagent parmi trois façons de mesurer la compréhension, mais ne le retiennent pas, préférant employer des tests de compréhension.

À notre connaissance, il faut attendre la période structuro-cognitiviste pour que ce critère soit utilisé par Richaudeau (1974), qui cherche à étudier ainsi certaines caractéristiques stylistiques et les performances en mémoire immédiate. Il prend en compte le nombre de mots correctement remémorés, une procédure d'évaluation qui est critiquée par Kintsch et ses collaborateurs (Kintsch *et al.*, 1975; Kintsch et Vipond, 1979; Miller et Kintsch, 1980), lesquels estiment qu'elle ne correspond à aucune réalité psychologique. Ces auteurs préfèrent employer le temps de lecture par proposition remémorée. On voit donc qu'une des difficultés de ce critère est la correction de la production : quelles unités faut-il privilégier ? N'y a-t-il pas dans le fragment lu des éléments plus significatifs que d'autres ?

Un second problème est justement l'influence des performances mémorielles sur la mesure de la compréhension. Richaudeau (1974, 85) rapporte un cas précis où un sujet n'ayant pas été très brillant en termes de rappel immédiat, était revenu le lendemain avec une version plus complète. Il est donc possible que des éléments soient compris et même intégrés en mémoire à long terme, mais ne soient pas entièrement récupérés au moment de l'évaluation. Enfin, à l'instar du temps de lecture, ce critère requiert de s'entretenir avec chacun des sujets pour chacun des textes du corpus. Cette solution n'est donc pas pratique et nous ne l'avons pas retenue.

4.2.7 Autres critères

À côté de ces quatre critères, d'autres possibilités ont parfois été envisagées dans la littérature, sans être vraiment utilisées. Nous les présentons brièvement dans cette section, en commençant par le résumé, qu'il soit écrit ou oral. Il s'agit, après lecture du texte cible, d'en

résumer les grandes lignes sans y avoir accès. Bien qu'il s'agisse d'une tâche de haut niveau qui requiert a priori une bonne compréhension du texte à synthétiser, Cornaire (1985, 56) dit que « les étudiants parviennent quelquefois à résumer un texte en l'ayant seulement compris de façon superficielle ». Par ailleurs, plus encore que dans le test de closure, cette tâche comporte un aspect de production, dans lequel les apprenants peuvent être plus ou moins performants, sans que cela soit lié à leur compréhension du texte.

Un autre aspect problématique est la méthode de correction. Kintsch et Vipond (1979, 338) insistent sur la difficulté de cette tâche, en particulier hors d'un cadre théorique précis. En effet, considérer le nombre de mots corrects ne rime à rien dès lors que la taille du fragment à restituer dépasse une ou deux phrases. Ce sont les idées qui doivent être prises en compte, ce qui nécessite un modèle de représentation du sens. Bien que certains modèles de ce type aient été proposés (Kintsch et van Dijk, 1978), ce critère n'a pas vraiment réussi à convaincre.

Un second critère, quant à lui utilisé à diverses reprises dans la littérature, consiste à étalonner les textes d'un corpus à l'aide d'une ou plusieurs formules de lisibilité existantes Citons comme exemple Forbes et Cottle (1953), qui annotent leurs textes en fonction de la moyenne des cinq formules suivantes : Dale et Chall (1948a), Flesch (1948), Lorge (1944a), Lewerenz et Yoakam. Quant à Farr *et al.* (1951), ils se limitent à la formule de Flesch. Cette approche est cependant justifiée par leur objectif, qui est de proposer une révision de la formule de Flesch, la plus proche possible des résultats originaux. Il est donc normal que leur critère soit l'objectif à atteindre. Cependant, de manière générale, on ne recommandera pas cette technique, vu son caractère circulaire.

Carver (1974a, 4-6) propose une technique assez proche du test de closure, qu'il appelle « reading input ». Plutôt que de supprimer un mot sur cinq, celui-ci est mélangé à des distracteurs parmi lesquels le lecteur doit le retrouver en s'aidant du contexte. Carver défend cette approche, car elle se rapproche plus de la lecture. En effet, un test de closure demande davantage de temps et de ressources mentales, puisqu'il est nécessaire d'effectuer diverses hypothèses concernant les réponses possibles, de sélectionner la meilleure, avant de prendre le temps de la coucher par écrit. Au contraire, le « reading input » nécessite moins d'efforts et se rapproche davantage d'une lecture normale.

Carver estime aussi que la correction de ce test est plus simple : elle peut être effectuée par un ordinateur via un scanneur. Il est toutefois nécessaire de rectifier les résultats obtenus afin de tenir compte des succès dus au hasard. Par ailleurs, la conception même du test est plus complexe : il s'agit de trouver de bons distracteurs. Dans une expérience, visant à étayer la validité de sa nouvelle approche Carver (1974a, 9-12) reprend les 36 passages de Miller et Coleman (1967) et les soumet à 180 étudiants sous la forme d'un test « reading-input ». Il obtient une corrélation de $0,87$ avec les résultats à des tests de closure sur ces mêmes textes.

Il conclut donc à la relative équivalence des deux procédures.

Une quatrième technique est proposée par Rothkopf (1980) : l' « empan de copie » ou « copying span » en anglais. Comme son nom l'indique, elle consiste à mesurer la longueur moyenne de la chaîne de mots qui peut être recopiée, par exemple dactylographiée, sans que le texte d'origine ne soit consulté à nouveau. L'emploi du « copying span » en lisibilité s'appuie sur l'hypothèse que les passages plus riches en information imposent davantage de contraintes sur la mémoire à court terme, laquelle est mise à contribution par ce type d'exercice. L'auteur effectue deux expériences où cinq dactylographes expérimentées doivent recopier 17 passages courts (de 83 à 159 mots), tandis que leur « empan de copie » et leur vitesse de lecture sont mesurés. Parallèlement, ces mêmes textes sont soumis à 120 étudiants sous la forme de divers tests de closure. Il en ressort (1) que l'ordonnancement des textes effectué à l'aide du « copying span » est cohérent ; (2) que cette technique est également fortement corrélée avec les résultats des tests de closure ($r = 0,749$ et $0,781$), (3) ainsi qu'avec les performances de la formule de Flesch (1948) ($r = 0,813$). On peut toutefois se demander dans quelle mesure la tâche requise ici nécessite de comprendre ce que l'on recopie !

Pour conclure ce point, citons encore rapidement quelques autres critères. Ainsi, Jacobson (1965) recourt, pour sa formule, au nombre de mots inconnus que contiennent les textes à étalonner, un test que cet auteur nomme le « Underlining Test ». Robert Shaw (1967, cité par Klare, 1974, 78) développe une formule en se basant sur le nombre d'erreurs commises par des élèves de la quatrième à la sixième année primaire sur un corpus de matériaux scientifiques, tandis que Smith et Senter (1967) s'appuient sur l'avis d'éditeurs. Quant à Dale et Tyler (1934), ils ont même envisagé de prendre en compte l'existence d'une réaction émotionnelle après la lecture comme le signe d'une bonne compréhension. Nous ne discuterons pas ici la validité de ces quatre derniers critères, que nous n'avons pas retenus pour notre étude.

4.2.8 Un choix pratique : les manuels scolaires

Pour clore notre discussion des différents critères qui s'offrent au chercheur en lisibilité, il convient de s'attarder sur une possibilité qui ne constitue pas un critère en soi, mais qui a néanmoins été largement employée pour la conception de formules de lisibilité. Elle consiste à se servir de textes tirés de manuels scolaires, dont le niveau a déjà été évalué par d'autres, que ce soit à l'aide d'une batterie de tests ou via un jugement humain.

Les plus connus des manuels ayant servi en lisibilité sont les leçons de McCall and Crabbs, que nous avons déjà présentées au Chapitre 2.2.3. Elles furent employées par certains concep-

teurs des formules les plus célèbres (Dale et Chall, 1948a; Flesch, 1948; Gunning, 1952; Kincaid *et al.*, 1975; Lorge, 1944a) et ont été étalonnées à partir de tests de compréhension. Cependant, il est rare que le contenu de manuels scolaires soit ainsi soumis à évaluation par une large population. Bien souvent, il est le reflet de l'avis d'un nombre réduit d'experts, lequel est ensuite confronté à l'usage des professeurs et peut être amélioré au fil des rééditions.

Se servir de matériaux scolaires dont le contenu est destiné à un niveau d'éducation spécifique est une idée qui n'est pas neuve. Dolch (1928) base déjà son étude sur des livres simplifiés pour enfants dont le niveau de difficulté est exprimé en niveaux scolaires américains. Il reprend ce même critère par la suite (Dolch, 1948) et le défend sur la base de l'argumentation suivante. Lorsqu'on dit qu'un texte est d'un niveau X de difficulté, on signifie habituellement qu'il est accessible à un étudiant de l'année scolaire X. Or, il existe des disparités importantes entre les élèves d'une même année, ce qui oblige à considérer cette indication comme une moyenne. Mais une moyenne de quoi ? Là aussi, d'une école à l'autre, on observe des disparités notables. Par conséquent, (Dolch, 1948, 230) en conclut :

> *Hence, if we cannot be sure just what children of any particular grade can read, we must conclude that the only conception of reading difficulty for any grade that is at present defensible is derived from the reading difficulty of the reading books provided for that grade.*

En avançant cet argument, Dolch s'éloigne de la définition classique de la lisibilité que nous avons présentée au Chapitre 2.1.1. Il ne s'agit plus de déterminer la difficulté de matériaux par rapport à une population de référence, mais de classer de nouveaux textes conformément aux normes en vigueur pour les manuels qui sont destinés à la population d'intérêt. Dans ce cas, le problème change de nature : de la création d'un outil psychométrique, on cherche à définir un classifieur de textes. Or, c'est justement l'optique qui est adoptée par la plupart des travaux en lisibilité computationnelle.

Cette vision particulière de la lisibilité est également épousée par Spache (1953). Il attribue aux livres de son corpus le niveau correspondant à l'année scolaire à laquelle ceux-ci sont destinés et démontre empiriquement l'intérêt d'une telle approche. Sa formule est en effet reconnue comme l'une des plus fiables pour les primaires ($r = 0,818$). Par la suite,Wheeler et Smith (1954) et Bloomer (1959) utilisent des *readers* (versions simplifiées de classiques dont le niveau est spécifié en fonction du système scolaire) comme critère, sans se justifier particulièrement.

Toutefois, c'est surtout aujourd'hui que cette seconde vision de la lisibilité fait fureur. Rappelons que depuis Si et Callan (2001), la conception d'une formule de lisibilité est vue

comme un problème de classification, ce qui s'accorde parfaitement avec la vision de Dolch (1948). Dès lors, il n'est pas étonnant que la majorité des chercheurs modernes recourent à des textes issus de ressources pédagogiques (Daoust *et al.*, 1996; Feng *et al.*, 2009; François, 2009a; Schwarm et Ostendorf, 2005; Tanaka-Ishii *et al.*, 2010) ou, simplement, annotés par d'autres experts humains (Collins-Thompson et Callan, 2004a; Heilman *et al.*, 2007; Si et Callan, 2001).

Si cette technique d'annotation est tellement populaire, c'est qu'elle présente plusieurs atouts. Tout d'abord, il s'agit indéniablement de la démarche la plus rapide pour obtenir un grand nombre de textes. Il suffit de sélectionner avec soin les manuels d'où seront tirés les textes du corpus, avant de scanner ceux-ci afin de disposer d'une version numérique. Aucun test sur sujet n'est requis, ce qui limite considérablement les coûts de constitution du corpus. Ensuite, comme Dolch (1948) l'a soutenu, échantillonner un large éventail de manuels produit une formule à la couverture plus large et qui ne risque pas d'être biaisée par les caractéristiques des sujets testés (issus d'une même ville, d'une même école, voire d'une même classe sociale). Enfin, une expérience d'Harrisson (1980) semble à nouveau valider empiriquement l'approche basée sur des textes scolaires. Il compare les prédictions de neuf formules pour 50 textes également évalués par des enseignants. Il en ressort que l'écart entre les formules célèbres (Dale et Chall, 1948a; Flesch, 1948; Mugford, 1970) et les jugements humains est inférieur à un *grade*. Il semblerait donc que les formules basées sur les leçons de McCalls and Crabbs ne soient pas si mauvaises que le prétend Stevens (1980).

Bien entendu, utiliser des textes extraits de manuels scolaires comporte aussi certains points faibles. Le principal d'entre eux est le risque que le corpus ainsi constitué soit sujet à une plus grande hétérogénéité. Dolch (1928) repère d'ailleurs, au sein de son corpus, une grande diversité entre des livres d'un même niveau scolaire en termes d'indices de lisibilité. Il opère le même constat par la suite (Dolch, 1948, 245), ajoutant que la cohérence des ouvrages des premières années est plus élevée que pour ceux des niveaux intermédiaires. Une des raisons qui expliquent cette hétérogénéité serait le manque de soin de certains éditeurs dans la sélection des matériaux destinés à un niveau. En tout cas, Anderson (1983) estime : « The use of publisher's grade levels as a criterion is suspect unless we know how publishers assign them and how accurate they are ».

4.2.9 Conclusion

Dans cette section, nous avons passé en revue l'essentiel des critères employés en lisibilité. Nous avons montré que la majorité des chercheurs dans le domaine se sont efforcés de trouver un critère qui mesure plus adéquatement la compréhension. Ainsi, les premiers concepteurs de formules, insatisfaits des jugements émis par les experts, se sont rabattus sur

les tests de compréhension. Ceux-ci ont toutefois assez vite révélé leurs faiblesses, de sorte que les chercheurs ont alors envisagé d'employer le test de closure à partir des années 60. Ce nouveau dispositif pédagogique n'a cependant pas réussi à faire l'unanimité et, à partir des années 70, on a observé un retour vers l'emploi des jugements humains, ainsi qu'une multiplication des critères employés.

Aujourd'hui, ce sont toujours les avis des annotateurs humains qui sont privilégiés dans le cadre de la lisibilité computationnelle. Les auteurs de ce paradigme ont opéré à la suite de Dolch (1948) un ajustement de la façon dont on envisage classiquement la lisibilité. Il ne s'agit plus de classer un ensemble de textes par rapport à une population dont on dispose d'un échantillon de taille réduite, mais d'étalonner des textes par rapport à un corpus scolaire faisant référence dans une matière ou une communauté. Ce faisant, l'objectif à atteindre se réduit à développer un classifieur textuel, qui court cependant le risque de n'être pas adapté à la réalité des pratiques si le corpus de référence ne l'est pas non plus. Par contre, il est facile, en adoptant cette démarche, de rassembler un nombre de textes suffisant pour exploiter des techniques d'apprentissage automatisé issues du TAL ou du domaine de la fouille de données.

Pour ces raisons, et parce qu'aucun critère ne s'est révélé optimal jusqu'à présent, nous avons finalement opté pour cette procédure d'annotation, largement employée ces dernières années. En effet, même si Chall (1958, 17) reproche aux formules découlant de ce critère de ne pas estimer la compréhension à la lecture, mais simplement la similarité de nouveaux textes avec le niveau de ceux en usage, cela correspond très précisément à l'optique de notre thèse : trouver automatiquement des textes d'un niveau similaire à ceux que le professeur de FLE emploie quotidiennement. Reste dès lors à identifier un ensemble de textes réservés à des lecteurs L2 et qui sont déjà annotés par des experts selon une échelle de difficulté pratique. Ce sont les aspects que couvre la section suivante.

4.3 Le corpus et sa constitution

4.3.1 Introduction : nos besoins

Au terme de l'examen critique des différents critères à notre disposition, nous avons choisi de rassembler des textes déjà annotés par des experts. Il a fallu réfléchir à un moyen pratique de constituer ce corpus, ainsi qu'identifier les besoins auxquels celui-ci devrait répondre. Ceux-ci sont triples et, bien qu'ils aient déjà été présentés de manière éparse, nous les reprenons ici :

Taille du corpus : Précédemment, nous avons clairement positionné notre travail au sein de

la lisibilité computationnelle et des techniques d'apprentissage automatisé. Par conséquent, notre corpus ne peut se réduire à quelques dizaines de textes, comme c'est le cas pour nombre de travaux précédents. La taille idéale est néanmoins difficile à spécifier, aussi avons-nous décidé de rassembler le plus de textes possible ;

Expression de la difficulté via une échelle pratique : Comme nous l'avons souligné à la fin du Chapitre 2.2.8, il nous semble que la récupération de la lisibilité par les études computationnelles a conduit à prêter une moins grande attention à l'aspect pratique des échelles de difficulté utilisées, du moins chez certains chercheurs. De notre côté, nous avons tenu à adopter une échelle pratique, c'est-à-dire une référence de la difficulté qui soit la plus évocatrice possible pour les utilisateurs en FLE (professeurs, apprenants, etc.). Par conséquent, un choix évident est l'échelle définie dans le CECR (Conseil de l'Europe, 2001) qui est aujourd'hui largement diffusée dans le milieu du FLE ;

Variation du contenu : La troisième contrainte dirigeant notre procédure de compilation est la nécessité de disposer de contenus et de type de textes variés. En effet, il y a fort à penser qu'une formule de lisibilité entraînée sur un type de texte particulier se généralise mal à d'autres types. Par ailleurs, il est intéressant de vérifier si des documents aussi divers qu'un poème, un extrait de roman, un dialogue, une publicité ou une recette de cuisine peuvent être évalués correctement à l'aide d'une même formule. C'est pourquoi nous avons pris soin de considérer différentes formes d'écrits dans notre corpus.

Confronté à ces besoins, une solution nous est rapidement apparue : extraire les textes dont nous avons besoin depuis des manuels de FLE publiés après 2001. Après cette date, le niveau des manuels est systématiquement indiqué en référence à l'échelle du CECR, qui est également celle employée quotidiennement par les professeurs pour situer leurs étudiants. Par conséquent, un outil capable de classer automatiquement de nouveaux textes en fonction des normes en vigueur dans les documents communément employés dans l'enseignement devrait se révéler particulièrement pratique.

Dans la suite de cette section, nous commençons par présenter brièvement le CECR et son échelle de difficulté (*cf.* Section 4.3.2). Ensuite, nous exposons les critères que nous avons employés pour sélectionner les manuels de FLE ayant servis à notre étude, ainsi que les critères retenus pour faire le tri parmi les textes de ces manuels (*cf.* Section 4.3.3). Le point suivant (*cf.* Section 4.3.4) liste les manuels retenus et détaille leur classification lorsque c'est nécessaire. Enfin, nous décrivons les caractéristiques du corpus obtenus par cette méthode, ainsi que les formats utilisés pour le stockage et l'annotation des textes (*cf.* 4.3.5).

4.3.2 Le Cadre européen commun de référence pour les langues

Circulant depuis 1996 sous la forme de brouillons, le CECR a été publié officiellement en 2001, en français et en anglais. Depuis lors, il a déjà été traduit dans plus de 20 langues européennes (Little, 2006, 167). Il constitue la dernière réalisation linguistique d'importance du Conseil de L'Europe, un organisme distinct de l'UE, qui a été fondé en 1949 pour « défendre les droits de l'homme, la démocratie parlementaire et la primauté du droit », ainsi que pour « promouvoir la conscience d'une identité européenne reposant sur des valeurs partagées et transcendant les différentes cultures » (Little, 2007, 646). Dans le cadre de ces différentes missions, le Conseil de l'Europe cherche depuis plusieurs années à valoriser le plurilinguisme au sein de l'UE, notamment au moyen du CECR, mais aussi, précédemment, via la publication du Niveau Seuil (Coste *et al.*, 1976) et la dissémination d'acquis théoriques tels que le développement de l'approche communicative ou la valorisation de l'autonomie des apprenants.

Les objectifs du CECR sont décrits dans le document lui-même (Conseil de l'Europe, 2001, 9) : il « offre une base commune pour l'élaboration de programmes de langues vivantes, de référentiels, d'examens, de **manuels**, etc. en Europe ». Il vise ainsi à faciliter la coopération entre les différents partenaires et institutions impliqués dans l'enseignement des langues étrangères, en fournissant une base commune qui permet une reconnaissance mutuelle des différentes certifications et veut aider les professeurs, mais aussi les concepteurs de cours ou de manuels, à situer leurs efforts par rapport à une référence. Dans ce but, le CECR a développé une échelle de compétences à deux dimensions : l'une verticale, l'autre horizontale.

L'échelle verticale, qui nous intéresse ici, utilise des descripteurs du type « est capable de » pour définir six niveaux de maîtrise communicationnelle : A1 (niveau introductif ou de survie) ; A2 (niveau intermédiaire) ; B1 (niveau seuil) ; B2 (niveau avancé ou utilisateur indépendant) ; C1 (niveau autonome ou de compétence opérationnelle effective) et C2 (maîtrise) (Conseil de l'Europe, 2001, 24-25). Le choix d'une échelle à six niveaux est justifié par les auteurs du CECR, car il existerait « un large consensus (encore que non universel) sur le nombre et la nature des niveaux appropriés pour l'organisation de l'apprentissage en langues [...] » (Conseil de l'Europe, 2001, 24). Notons toutefois que, d'après Little (2006, 168), elle a été mise au point dans le cadre d'un projet de recherche suisse, qui suggérait aussi l'adoption de 9 niveaux : A1, A2, A2+, B1, B1+, B2, B2+, C1, C2.

À côté de cette échelle verticale, un axe horizontal décrit la possibilité d'évoluer en « élargissant ses aptitudes plutôt qu'en les accroissant dans une même catégorie » (Conseil de l'Europe, 2001, 20), c'est-à-dire en intégrant de nouveaux usages communicationnels du langage (lecture, interaction, expression orale, etc.). Figueras (2007, 673) souligne que la

réponse du public a été nettement plus positive envers l'échelle verticale qu'envers l'échelle horizontale. Celle-ci a rapidement été adoptée par les concepteurs de tests standardisés et de manuels, mais aussi dans les écoles de langues. Deux questions se posent cependant concernant cette échelle que nous avons également décidé d'adopter dans cette thèse : (1) quels sont les critères employés pour situer les différentes compétences du CECR à un niveau plutôt qu'à un autre ? (2) Quel est le crédit que l'on peut accorder à des manuels ou des tests qui effectuent eux-mêmes le lien entre leurs contenus et les descripteurs de CECR ?

La réponse à notre première question est apportée dans l'Annexe A du CECR, où sont décrites les techniques employées pour étalonner les différents descripteurs. Il s'agit en réalité d'une méthodologie mixte, qui combine l'avis d'experts avec des études qualitatives et quantitatives « dans un processus cumulatif de complémentarité » (Conseil de l'Europe, 2001, 150). Au départ, un groupe d'experts se base sur une série d'échelles de compétences déjà existantes pour produire un fonds initial de descripteurs. Ensuite, ces descripteurs sont répartis dans les différentes catégories qu'ils cherchent à décrire et sont triés en fonction de leur difficulté. Cette étape qualitative est réalisée via 32 ateliers de travail réunissant des professeurs de langues. Ensuite, des apprenants sont testés afin de vérifier si leurs compétences correspondent aux descripteurs retenus, cette adéquation étant rapportée sur une échelle de 0 à 4 par les professeurs. Cette étape qualitative est confirmée et affinée à l'aide d'une approche quantitative qui recourt au modèle de Rasche (1961).

En résumé, il ressort que l'étalonnage du CECR a été réalisé avec soin et n'a pas à rougir en comparaison des autres échelles en vigueur dans l'enseignement des langues. Plus délicat est le second problème. En effet, si l'engouement des concepteurs de manuels envers le CECR est réel et si la quasi-totalité des ouvrages publiés après 2001 porte une mention de niveau exprimée selon l'échelle européenne, rien ne nous certifie que ce label est représentatif de la difficulté réelle du manuel. Alderson (2007, 661) déplore d'ailleurs que le Conseil de l'Europe n'ait pas mis au point, pour les manuels ou les tests standardisés, un système de validation du respect de l'échelle, comme c'est le cas pour les portfolios [5]. Ce manque de contrôle est susceptible d'entraîner une hétérogénéité du contenu des manuels soi-disant situés à un même niveau. Cependant, nous postulons que cet effet sera limité, dans une certaine mesure, par le fait que les manuels sont réellement employés par des apprenants et des professeurs, lesquels sont susceptibles d'effectuer un retour sur les matériaux des manuels lorsque ceux-ci ne conviennent pas. Par conséquent, ce manque de standardisation n'a pas été retenu comme suffisant pour invalider notre approche.

5. Un portfolio européen des langues est un document dans lequel un apprenant peut consigner ses connaissances linguistiques et culturelles et qui vise à le faire réfléchir sur son apprentissage. Il offre donc une sorte de diagnostic exprimé en fonction de l'échelle du CECR.

4.3.3 Les critères de sélection

Notre description de la méthodologie utilisée pour définir l'échelle du CECR ayant souligné sa validité, nous poursuivons ici notre réflexion sur la constitution d'un corpus de textes étalonnés en fonction de cette échelle. Il ressort qu'il est possible, bien que potentiellement problématique, d'employer les manuels de FLE comme source de textes. Dès lors, le niveau de ces textes est établi d'après l'hypothèse suivante :

Hypothèse 4.1 *Le niveau d'un texte sera considéré comme équivalent à celui du manuel dont il est tiré.*

Cette hypothèse peut évidemment être discutée et le sera d'ailleurs (*cf.* Section 4.4). Elle présente toutefois l'aspect pratique requis pour répondre aux besoins identifiés précédemment (*cf.* Section 4.3.1).

Pour autant, il n'est pas souhaitable d'employer l'ensemble des ouvrages publiés pour le FLE, ni de retenir tous les textes qu'ils contiennent. Il existe en effet des publications qui ciblent un type de public particulier. On rencontre des livres qui adaptent leurs contenus pour les jeunes enfants et d'autres qui se concentrent sur des apprenants ayant une langue maternelle particulière, dont les caractéristiques influencent l'étalonnement des matériaux. Citons également les manuels de français sur objectifs spécifiques (FOS), qui visent à faire acquérir des genres de discours propres à certains domaines, tels que les affaires, le tourisme, la médecine, etc., et dont la planification est dirigée par les besoins pragmatiques des apprenants (Richer, 2008).

Or, comme nous l'avons expliqué au début du Chapitre 2.1.1, une formule de lisibilité se doit de se limiter à une population particulière, dont les caractéristiques sont considérées comme suffisamment homogènes pour être négligées dans le modèle. Il n'est donc pas question de mélanger à la fois des adultes et des enfants, des apprenants étudiant le tourisme ou les affaires, etc. Aussi, avons-nous défini un ensemble de critères de sélection dont le choix a été motivé par la volonté de conserver un public d'apprenants homogène pour l'intégralité du processus d'échantillonnage et, par conséquent, une population homogène de textes. Ces critères s'organisent en deux familles : ceux qui régissent le choix des manuels et ceux qui gouvernent la sélection des textes à l'intérieur de ceux-ci.

La sélection des manuels : Pour les raisons évoquées ci-dessus, nous n'avons retenu pour notre étude que des manuels de FLE remplissant les conditions suivantes :

1. **Compatibilité avec le CECR** : ayant opté pour l'échelle du CECR pour cette étude, il faut évidemment que le niveau des manuels pris en compte soit exprimé en fonction de celle-ci. C'est généralement le cas des ouvrages postérieurs à 2001 ;

2. **Modernité** : les ouvrages doivent être récents, afin de s'assurer que le langage et les réalités culturelles qui seront modélisés correspondent bien à la manière actuelle d'enseigner et de parler le FLE. En effet, les manuels antérieurs à 2001 font, par exemple, référence au franc et non à l'euro. De même, les entités nommées et les sujets abordés évoluent aussi en fonction de la date de publication. Soulignons cependant que ce second critère n'ajoute guère de restrictions par rapport à la condition 1, laquelle a déjà éliminé la majorité des manuels antérieurs à 2001 ;
3. **Type de public** : le type de public ciblé par les manuels est par contre un aspect crucial, afin de respecter une cohérence dans l'échantillonnage. Nous avons décidé de restreindre notre formule à une population d'adultes ou de jeunes adultes. Dès lors, seuls les manuels dédiés à ce public particulier ont été sélectionnés ;
4. **Type de lecture** : d'après l'hypothèse 2.3 posée dans l'introduction à cette seconde partie de la thèse, nous devons postuler que le type de lecture visé par notre formule correspond à celle enseignée dans une classe visant un apprentissage généraliste du FLE. C'est pourquoi nous avons rejeté l'emploi de manuels de FOS ;
5. **Prise en compte du background des apprenants** : la dernière condition relève de la prise en compte de la langue maternelle des apprenants. Les manuels ciblant des apprenants d'une L1 particulière ont tendance à employer celle-ci pour l'exposition de consignes ou de règles linguistiques, et surtout à adapter leur contenu en fonction des proximités ou des interférences entre cette L1 et la L2 étudiée. Or, notre formule n'étant actuellement pas pensée pour une paire de langues déterminées, nous n'avons pas non plus utilisé ce type de manuels.

La sélection des textes : Sur la base de ces cinq critères, 28 manuels ont été retenus (*cf.* Section 4.3.4 pour le détail). Nous les avons ensuite parcourus à la recherche de textes appropriés pour notre corpus. Là aussi, il a été nécessaire de lister des critères de sélection afin de ne retenir que les textes appropriés à nos objectifs, à savoir :

1. **Type de tâche** : l'objectif d'une formule de lisibilité étant d'estimer la difficulté de matériaux à la lecture, il importe de ne sélectionner que des textes associés à une tâche de compréhension écrite. En effet, les textes présents dans les manuels sont habituellement associés à des consignes et c'est la combinaison des deux qui engendre la difficulté de l'exercice. Ainsi, nous avons par exemple rejeté les transcriptions des exercices de compréhension orale, car il nous est apparu que la complexité du langage qui y est utilisé est inférieure à celle d'un texte réservé à une tâche de compréhension écrite. Par ailleurs, en ce qui concerne les niveaux A1 et A2 où l'on rencontre de très brefs fragments, tout ce qui ne constitue pas au moins une phrase complète a été éliminé. Cela comprend des listes d'éléments, des exercices à trous et autres exercices à lacunes ;

2. **Sélection des informations** : eu égard à la nécessité d'exporter le textes des documents en vue d'un traitement TAL subséquent, nous n'avons pas conservé les éléments graphiques tels que les images, tableaux, graphiques, photos et autres illustrations ;
3. **Les consignes** : la plupart des textes sont accompagnés de quelques lignes qui introduisent les consignes à suivre pour cette activité. Bien que la langue employée pour ces instructions soit généralement contrôlée, sa difficulté ne s'accorde pas nécessairement avec celle du texte. Elles risquent d'être plus complexes que celui-ci dans les manuels pour débutants et nettement plus simples dans les niveaux avancés. De plus, elles comportent un aspect répétitif peu intéressant à modéliser. Enfin, mesurer la difficulté combinée des consignes et des textes dépasse tout simplement les objectifs de cette étude. Pour ces différentes raisons, nous avons donc systématiquement ignoré ce matériau linguistique ;
4. **Taille des textes** : à nouveau, dans les manuels pour débutants, il est possible de trouver des exercices ne comportant que quelques phrases très brèves. Afin de conserver une certaine cohérence dans le contenu et le style, il nous a semblé préférable de ne pas considérer des phrases isolées, à moins qu'elles ne participent d'une même activité et soient en nombre suffisant ;
5. **Genre textuel** : par contre, nous n'avons pas été restrictif en ce qui concerne les genres textuels pris en considération. Ainsi, nous avons rassemblé les neufs types de textes suivants : des dialogues, des lettres, des courriels, des publicités, des chansons, des poèmes, des recettes de cuisine, des suites de phrases déconnectées et des textes. Le genre a cependant fait l'objet d'une annotation (*cf.* Section 4.3.5) afin de pouvoir revenir sur ce choix par la suite.

4.3.4 Les manuels retenus et leur classication

En employant ces différents critères, un corpus 2042 textes, comportant plus de 530 000 tokens, a été constitué. Restait dès lors à attribuer à chaque document un niveau de difficulté en accord avec l'échelle du CECR. Sachant que le manuel dont chacun d'eux a été extrait est lui-même situé à un niveau donné et que ce même niveau est attribué au texte en fonction de l'hypothèse du point 4.3.3, cette tâche apparaît facile à première vue. C'est sans compter une difficulté inattendue. En effet, certains manuels couvrent plus d'un niveau et présentent dès lors une étiquette mixte, par exemple A1/A2. Dans ce cas, comment distinguer les textes du niveau A1 de ceux qui doivent être annotés A2 ?

Cette problématique est encore accentuée par un autre fait lié à notre choix d'échelle. En effet, parallèlement à l'échelle à 6 degrés du CECR, nous avons estimé intéressant de subdiviser certains des niveaux, à savoir le A1, A2 et B1, en deux sous-niveaux (A1 et A1+ ; A2

et A2+ et B1 et B1+). Ce système en arborescence, préconisé dans le CECR même (Conseil de l'Europe, 2001, 31), vise à mieux modéliser les rapides changements linguistiques observables dans les textes des premiers niveaux de notre corpus [6]. Il rejoint par ailleurs en partie le point de vue du projet de recherche suisse à l'origine du CECR (Little, 2006, 168), dans lequel 9 niveaux étaient décrits.

Par conséquent, dans la suite de cette section, nous décrivons comment les contenus de chaque manuel retenu ont été répartis en fonction des 9 niveaux de notre échelle [7]. Les critères de répartition varient d'un manuel à l'autre et reposent essentiellement sur les informations que nous avons pu collecter dans les introductions pédagogiques de chaque ouvrage. Celles-ci permettent, dans la plupart des cas, d'identifier plus précisément les sous-niveaux couverts. De manière générale, lorsqu'un ouvrage dit explicitement couvrir de façon équivalente deux niveaux sans plus d'information, nous avons considéré la première moitié des leçons comme ressortissant au premier niveau et la seconde moitié au second. Pour le reste, la répartition propre à chaque manuel est décrite dans la description suivante :

Alter Ego 1 (Berthet *et al.*, 2006a) : Cet ouvrage, conçu par des professeurs-formateurs à l'Alliance Française, « s'adresse à des débutants et vise l'acquisition des compétences décrites dans les niveaux A1 et A2 (en partie) » du CECR (p.3). Bien que sur la couverture extérieure, seul le niveau A1 soit spécifié, une indication « A1 > A2 » apparaît à partir du cinquième dossier et se change en « A1/A2 » à partir du septième. Nous avons dès lors classifié les dossiers comme suit :

A1 : dossiers 1 à 3

A1+ : dossiers 4 à 6

A2 : dossiers 7 à 9

Connexions 1 (Loiseau et Mérieux, 2004) : ce livre s'adresse à des débutants et semble se limiter au niveau A1. Nous l'avons dès lors réparti équitablement :

A1 : unités 1 à 6

A1+ : unités 7 à 12

Festival 1 (Poisson-Quinton *et al.*, 2005a) : ce manuel comporte comme seule indication de niveau : le sigle A1 sur la couverture. Nous l'avons dès lors considéré ainsi :

A1 : unités 1 à 3

A1+ : unités 4 à 6

Ici 1 (Abry *et al.*, 2008b) : « prépare à la certification du niveau élémentaire A1 » (p.2). Ses textes ont été répartis de la façon suivante :

A1 : unités 1 à 3

6. Une autre façon de traiter l'évolution interne à un niveau aurait consisté à répartir les textes selon une échelle continue et non pas une classification. Nous ne l'avons pas considérée, car cela aurait probablement accentué les problèmes d'hétérogénité des manuels dont nous parlerons par la suite, tout en produisant des résultats non directement compatibles avec l'échelle du CECR.

7. Le système à 6 niveaux n'est pas décrit explicitement, puisqu'il découle logiquement du rassemblement des sous-niveaux.

A1+ : unités 4 à 6

Panorama 1 (Girardet et Cridlig, 1996a) : cette série est la seule à être antérieure à 2001. Nous l'avons cependant utilisée, car elle a fait l'objet d'une adaptation aux normes du CECR. D'après le catalogue 2008 de CLE International, cet ouvrage recouvrerait les niveaux A1 et A2 (en partie). Nous l'avons dès lors classifié comme suit :

A1 : unités 1 et 2

A1+ : unités 3 et 4

A2 : unités 5 et 6

Rond Point 1 (Labascoule *et al.*, 2005) : Ce manuel assez particulier est l'un des premiers qui a fait reposer son approche sur la perspective actionnelle. Il est présenté par ses auteurs comme couvrant les niveaux A1 et A2. Son étalonnage a été difficile : les trois premières unités ne comportent à peu près aucun texte et la plupart d'entre eux sont uniquement destinés à une compréhension globale (exemple, p.15) et, par conséquent, sont plus complexes. Nous avons finalement établi la classification suivante :

A1 : textes des pages 17, 19 et 22-23.

A1+ : textes des pages 15 et 24 à 54

A2 : unités 6 et 7

A2+ : unités 8 et 9

Taxi ! 1 (Capelle et Menand, 2003) : « La méthode permet à l'apprenant de se préparer aux épreuves du DELF 1er degré unité A1. » (p.3). Nous avons donc simplement divisé le contenu en deux :

A1 : unités 1 à 5

A1+ : unités 6 à 9

Tout va bien ! 1 (Augé *et al.*, 2005a) : Cette méthode couvre les niveaux A1 et A2 (en partie). D'après le catalogue 2008 de CLE International, il s'agirait d'un manuel pour faux-débutants, adaptés à un public de langues romanes et qui suit un rythme d'apprentissage soutenu. Nous avons dès lors réparti les contenus comme suit :

A1+ : unités 1 à 6

A2 : unités 7 à 12

Connexions : Préparation au DELF 1 : ces textes ont été récupérés en ligne sur le site de l'éditeur Didier [8]. Ils visent à préparer le DELF A1 et nous les avons donc classés comme A1+.

Alter Ego 2 (Berthet *et al.*, 2006b) : Ce manuel, réalisé par la même équipe de professeurs-formateurs de l'Alliance Française qu'*Alter Ego 1*, « vise l'acquisition des compétences décrites dans les niveaux A2 et B1 (en partie) » du CECR (p.3). On retrouve également des indications en début de dossiers qui nous ont conduit à diviser les leçons de la manière suivante :

A2 : dossiers 1 à 3

A2+ : dossiers 4 à 6

B1 : dossiers 7 à 9

Compréhension écrite 2 (Poisson-Quinton, 2005) : ce manuel « correspond aux niveaux A2 et A2+ du Cadre européen » (2005, 3). Il a donc été étalonné de la façon suivante :

8. Récupérés en 2008 à l'adresse suivante : http ://www.didierconnexions.com/zip/connexions1.zip.

A2 : leçons 1 à 6

A2+ : leçons 7 à 15

Ici 2 (Abry *et al.*, 2008a) : « prépare à la certification du niveau élémentaire A2 » (p.2). Ses textes ont été répartis de la façon suivante :

A2 : unités 1 à 3

A2+ : unités 4 à 6

Panorama 2 (Girardet et Cridlig, 1996b) : d'après le catalogue 2008 de CLE International, il recouvrerait les niveaux A2 et B1 (en partie). Nous l'avons dès lors classifié ainsi :

A2 : unité 1

A2+ : unités 2 à 4

B1 : unités 5 et 6

Taxi ! 2 (Menand *et al.*, 2003) : La méthode « permet à l'apprenant de se préparer à l'épreuve A2 du DELF 1er degré. » (p.3). Comme pour le manuel A1, nous avons divisé en deux les leçons :

A2 : unités 1 à 5

A2+ : unités 6 à 9

Tout va bien ! 2 (Augé *et al.*, 2005b) : « *Tout va bien ! 2* vise l'acquisition d'une compétence de communication d'un niveau intermédiaire. les objectifs et contenus de la méthode ont été déterminés à partir des niveaux A2 et B1 » du CECR (p.5). De plus, il s'adresse à des apprenants ayant déjà abordé le niveau A2. C'est pourquoi, nous catégorisé les leçons de la manière suivante :

A2+ : unités 1 à 4

B1 : unités 5 à 6

Connexions : Préparation au DELF 2 : comme pour ceux du niveau A1, ces textes ont été récupérés en ligne sur le site de l'éditeur Didier [9]. Puisqu'il s'agit d'une préparation au DELF A2, nous les avons considérés comme A2+.

Activités pour le CECR B1 (Parizet *et al.*, 2006) : les contenus de ce manuel sont regroupés par compétences (compréhension écrite, production écrite et interaction écrite) à propos desquelles les auteurs disent : « elles correspondent aux niveaux B1 et B1+ et sont donc de difficulté différente dans la mesure où l'écart est important entre le début de B1 et B1+ » (p.3). Par conséquent, nous avons postulé l'existence d'une progression au sein de chacun des compétences :

B1 : p.13 à p.40 et p.70 à p.114

B1+ : p.42 à p.54 et p.116 à p.168

Par ailleurs, il faut noter que certains textes retenus étaient, exceptionnellement, associés à une tâche de production ou d'interaction orale, laquelle nécessitait toutefois la lecture attentive du texte retenu avant d'être effectuée.

Alter Ego 3 (Dollez et Pons, 2006) : il « s'adresse à des apprenants ayant acquis un niveau A2. Il vise l'acquisition des compétences décrites dans le niveau B1 » du CECR (p.3). Sur les neufs dossiers, les deux derniers sont indiqués « B1 > B2 ». Cette fois, le manuel ne déborde pas sur le niveau B2, aussi avons-nous réparti les leçons comme suit :

9. Récupérés en 2008 à l'adresse suivante : http ://www.didierconnexions.com/zip/connexions2.zip.

B1 : dossiers 1 à 5

B1+ : dossiers 6 à 9

Compréhension écrite - Niveau 3 B1/B1+ (Poisson-Quinton et Mimran, 2007) : cet ouvrage « correspond aux niveaux B1 et B1+ du Cadre européen » (p.3). Par conséquent, son contenu a été subdivisé comme suit :

B1 : unités 1 à 6

B1+ : unités 7 à 15

Festival 3 (Poisson-Quinton *et al.*, 2005b) : il s'annonce comme couvrant les niveaux B1 et B1+ (Poisson-Quinton *et al.*, 2005b, 3). Nous l'avons dès lors segmenté en deux :

B1 : unités 1 à 3

B1+ : unités 4 à 6

Panorama 3 (Frerot et Girardet, 1997) : d'après le catalogue 2008 de CLE International, il recouvrerait les niveaux B1 et B2 (en partie). Nous l'avons dès lors classifié ainsi :

B1 : unités 1 et 2

B1+ : unités 3 à 8

B2 : unités 9 à 12

Rond Point 2 (Flumian *et al.*, 2004) : Ce manuel déclare couvrir le contenu du niveau B1. Nous avons donc simplement divisé ses unités en deux :

B1 : unités 1 à 5

B1+ : unités 6 à 9

Taxi ! 3 (Johnson *et al.*, 2004) : « il suit les recommandations du niveau B1 » du CECR (p.3). Nous avons divisé les unités en deux :

B1 : unités 1 à 6

B1+ : unités 7 à 12

Tout va bien ! 3 (Augé *et al.*, 2006) : « les objectifs et contenus de la méthode ont été déterminés à partir du niveau B1 » du CECR. Comme nous avons déjà attribué à B1 la fin de *Tout va bien ! 2*, les textes sont ici répartis selon un ratio 1/3 et 2/3 :

B1 : unités 1 et 2

B1+ : unités 3 à 5

Activités pour le CECR, Niveau B2 (Corsain *et al.*, 2007) : comme annoncé dans le titre, cet ouvrage couvre le niveau B2 et il est structuré par compétences. Les textes retenus sont associés à une tâche de compréhension orale ou à une tâche de production orale précédée d'une lecture et ils ont été considérés comme B2.

DELF B2, 200 activités (Bloomfield et Daill, 2006) : l'ensemble des textes de ce manuel sont classés B2.

Rond Point 3 (Capucho *et al.*, 2007) : idem, pour ce titre.

Activités pour le CECR, Niveau C1/C2 (Kober-Kleinert *et al.*, 2007) : ce manuel couvre les niveaux C1 et C2 et est organisé par compétences. Il mêle les activités C1 et C2, dont le niveau est heureusement signalé à chaque fois. C'est sur base de cette information que nous avons conçu notre classement, lequel n'est pas repris ici en détail.

DALF C1/C2, 250 activités - Le Nouvel Entraînez-vous (Lescure *et al.*, 2007) : cet ouvrage est également organisé par compétences et divisé en une section C1 et une autre C2. Dès lors, les textes n'ont pas été difficiles à classer entre les deux niveaux. :

C1 : textes issus des pages 64 à 178

C2 : textes issus des pages 205 à 252

Réussir le DALF C1/C2 (Chevallier-Wixler *et al.*, 2007) : ce manuel, organisé par compétences, présente l'avantage non négligeable d'être divisé en une partie C1 et l'autre C2. La classification est donc facile :

C1 : textes issus des pages 37 à 127

C2 : textes issus des pages 142 à 184

Après que la sélection et l'étalonnage des textes ont été opérés, nous nous sommes occupé de numériser chacun d'eux. Cette opération été réalisée à l'aide d'un programme d'OCR dont les résultats ont ensuite été révisés manuellement [10], afin de réduire au maximum les erreurs dues à ce processus. Une fois cette opération terminée, nous avons obtenu un corpus dont la répartition par série de manuels est synthétisée dans le Tableau 4.1. Nous remarquons, à la lecture de ce tableau, que le nombre de textes par niveau est – malheureusement – assez inégal. Ce problème vient du fait qu'il existe peu de manuels spécifiques pour les niveaux C1 et C2, probablement parce que comme les apprenants avancés sont capables de comprendre la majorité des écrits en français, les professeurs peuvent dès lors se passer plus aisément d'ouvrages pédagogiques. Cette constatation aura des répercussions sur notre modèle de difficulté, dont nous reparlerons par la suite 4.4.2.

	A1	A1+	A2	A2+	B1	B1+	B2	C1	C2
Activités CECR	/	/	/	/	41	39	50	63	8
Alter Ego	46	44	61	31	74	42	/	/	/
Comp. écrite	/	/	34	53	39	50	/	/	/
Connexions	34	26	/	/	/	/	/	/	/
Connexions : prep. DELF	/	11	/	12	/	/	/	/	/
Delf/Dalf	/	/	/	/	/	/	31	78	19
Festival	42	34	/	/	28	26	/	/	/
Ici	13	28	25	17	/	/	/	/	/
Panorama	31	27	50	48	56	57	41	/	/
Rond-point	3	19	4	7	21	19	76	/	/
Réussir Dalf	/	17	/	/	/	/	/	43	22
Taxi !	27	/	23	21	56	51	/	/	/
Tout va bien !	/	50	36	56	45	37	/	/	/
Total	196	256	233	245	360	321	198	184	49

Tableau 4.1 – Nombre de textes par niveau, pour chaque série de manuels pris en compte

Au terme de ce long travail de sélection des manuels et de numérisation des textes, nous

10. À ce propos, nous remercions ici vivement Bernadette Dehottay pour son aide précieuse en ce qui concerne la numérisation des textes et leur vérification manuelle.

avons pu entamer la dernière étape de constitution de notre corpus, à savoir définir un format d'annotation et de sauvegarde des données.

4.3.5 Caractéristiques du corpus

Sauvegarde des fichiers

Une caractéristique importante d'un corpus est la manière dont il est stocké informatiquement. Celle-ci a une influence sur les opérations postérieures effectuées et détermine la facilité d'accès aux données. Pour la sauvegarde des textes, nous avons envisagé deux architectures différentes. La première agrège, en un seul fichier texte, tous les documents dont le début et la fin sont délimités à l'aide de balises spécifiques. À l'inverse, la seconde utilise un fichier par texte et organise le corpus sous la forme d'une hiérarchie de répertoires. Le répertoire parent contient l'ensemble du corpus ; les sous-répertoires fils correspondent aux 6 ou 9 niveaux du CECR et chacun d'eux comprend les manuels ou parties de manuels qu'il recouvre.

Le fichier unique présente l'avantage de limiter les ouvertures fichiers lors de la consultation et d'éviter l'implémentation d'une fonction spécifique nécessaire pour explorer l'architecture des répertoires et récupérer le chemin de chaque fichier. Cette solution apparaît donc meilleure à première vue. Cependant, elle rend l'accès manuel aux données malaisé. En effet, l'ouverture d'un fichier de cette taille est toujours susceptible d'entraîner un ralentissement de la machine et, une fois ouvert, l'utilisateur doit encore réussir à retrouver le texte qu'il cherche.

À l'inverse, la seconde solution autorise une exploration manuelle plus aisée, puisque la hiérarchie de fichiers respecte la logique du classement que nous avons présenté au point 4.3.4. Par ailleurs, elle est nettement plus efficace lorsqu'on ne désire traiter qu'une partie du corpus – par exemple, pour effectuer des tests –, puisqu'il n'est alors pas nécessaire de parser l'ensemble du corpus. C'est pourquoi nous avons finalement opté pour la deuxième solution.

Dès lors, les textes de chaque manuel sont sauvegardés dans le répertoire correspondant, sous un nom uniquement constitué de la page d'où vient le texte, suivi de l'extension .txt (ex. : p98.txt). Il est cependant possible d'obtenir, à partir du chemin complet du fichier, le niveau et le nom du manuel d'où est extrait le texte (ex. : Corpus9/a2+/Taxi2/p98.txt). Nous n'avons pas rencontré de difficulté particulière avec cette architecture au cours de nos manipulations du corpus. Il serait toutefois possible de recourir à une base de données de type MySQL, afin d'obtenir un résultat assez similaire et probablement plus aisément

manipulable.

Format d'annotation

Pour chacun des textes retenus, nous disposons des informations suivantes : son niveau, son genre textuel et son titre, ainsi que le nom, la leçon et la page du manuel dont il est extrait. En TAL, lorsqu'on veut introduire des méta-informations au sein d'un texte, il est commun de recourir à des formats d'annotations balisés tels que HTML, XML (Ide, 1998), voire des formats plus spécifiques comme le TEI [11] (Ide et Veronis, 1995; Sperberg-McQueen *et al.*, 1994) ou XCES (Ide *et al.*, 2000).

Cependant, ces formalismes sont relativement complexes et, bien qu'ils présentent de nombreux avantages, ils nous ont semblé trop lourds à mettre en oeuvre pour une tâche aussi simple qu'ajouter 6 méta-informations au début de chaque fichier texte. Nous avons donc défini notre propre système de balises, qui respectent un format simple de type [attribut] : [valeur]. Leurs noms et valeurs sont détaillés dans la liste suivante :

Niveau : un niveau parmi les 6 suivants : A1, A2, B1, B2, C1 ou C2
Seuls les 6 niveaux de base du CECR sont utilisés pour l'annotation. En réalité, le niveau du texte varie d'un corpus à l'autre (selon qu'il y ait 6 ou 9 classes) et est récupéré dynamiquement à partir du chemin du fichier ;

Unité : [numéro de l'unité dont est issu le texte]/[nombre total d'unités pour le manuel]
Ce format fractionnel nous a paru propice dans le cas où nous désirerions subdiviser les niveaux en fonction des unités pour obtenir une difficulté exprimée selon une échelle continue ;

Source : nom du manuel dont est extrait le texte ;

Page : page(s) dont est extrait le texte ;

Date : date de publication du manuel. Cela ne correspond donc pas au moment de rédaction du texte, mais à sa publication dans un contexte FLE ;

Type_Texte : le genre de texte. Ce champ peut prendre les valeurs suivantes : Texte, Phrases, Dialogue (y compris les interviews), Lettre, Mail, Publicité (y compris les prospectus, petites annonces, etc.), Poème et Recette. La classification a été réalisée manuellement et n'a pas été difficile à mener à bien, étant donné que la majorité de ces genres textuels sont identifiables par des indices spécifiques ;

Titre : le titre du texte.

11. Le document reprenant l'ensemble des principes directeurs du TEI est hébergé à l'adresse suivante : `http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/`.

Après ces sept lignes, chacun des fichiers du corpus contient une ligne vide qui sépare celles-ci du corps du texte. Ce dernier a également fait l'objet de certains traitements manuels lors de la numérisation, afin de gérer certaines particularités de la mise en page. Nous les décrivons dans le point suivant.

Les formats des textes

Si numériser un extrait de roman ne nécessite pas d'adaptation particulière de la mise en page, il en va autrement des mails, des dialogues de théâtre, ou pire, des prospectus publicitaires. Nous avons ainsi établi les normes suivantes, qui ont été appliquées par les deux personnes en charge de la numérisation et de sa vérification.

Texte : il s'agit du format standard. Chaque ligne du fichier correspond à un paragraphe. Les mots découpés par un retour à la ligne ont été reconstitués afin d'éviter des problèmes de reconnaissance automatique. De même, les notes de bas de page ont été supprimées ;

Phrases : une phrase par ligne de fichier. Les correcteurs ont pris soin de vérifier qu'elles se terminent toujours par une ponctuation ;

Dialogue : chaque nouvelle réplique correspond à un retour à la ligne. Celles-ci peuvent être introduites soit par un tiret, soit par le nom du personnage qui parle, en lettres capitales ;

Lettre, Mail : la première ligne du texte est réservée à l'en-tête. Ensuite, le corps du texte suit le format standard spécifié ci-dessus et les deux dernières lignes sont respectivement réservées aux formules de politesse et à la signature ;

Publicité : les formats rencontrés étant si divers, nous avons cherché à respecter le plus possible la mise en page originelle tout en nous accommodant des limitations des fichiers au format *.txt* ;

Poème, Chanson : une ligne par vers. La ponctuation d'origine (ou l'absence de ponctuation) est respectée ;

Recette de cuisine : les informations de temps, de cuisson et les ingrédients sont placés sur une ligne différente. Pour ce qui est du corps de la recette, il suit la norme standard (Texte).

Quelques chiffres

Pour conclure cette description de notre corpus, les tableaux 4.2 et 4.3 présentent quelques chiffres offrant une idée de la taille du corpus et de sa diversité. Le tableau 4.2 décrit la

distribution bivariée des textes (en nombre de textes et de tokens) en fonction des niveaux du CECR et des genres textuels décrits ci-dessus. Le tableau 4.3 reprend ces mêmes données, mais ramenées à des pourcentages.

	A1	A1+	A2	A2+	B1
Chanson	/	2(234)	1(549)	/	/
Dialogue	81(8967)	72(14309)	41(9559)	31(8431)	24(5408)
E-mail	10(1128)	7(1170)	4(469)	2(190)	10(1281)
Lettre	12(891)	12(1358)	7(908)	11(1301)	26(8225)
Poème	/	7(896)	6(977)	4(521)	2(366)
Phrases	27(3156)	29(3916)	13(2647)	8(1483)	8(1165)
Publicité	6(766)	14(1891)	15(1173)	6(797)	10(2788)
Recette	2(203)	/	3(304)	1(118)	/
Texte	58(7456)	113(16251)	143(25789)	182(39901)	280(71438)
Total	196(22567)	256(40025)	233(42375)	245(52742)	360(90671)

	B1+	B2	C1	C2	Total
Chanson	/	/	/	/	3(783)
Dialogue	15(5732)	5(1698)	/	/	269(54104)
E-mail	/	6(1381)	/	/	39(5619)
Lettre	8(1687)	12(2812)	8(2144)	1(398)	97(19724)
Poème	7(939)	12(1469)	1(272)	/	39(5440)
Phrases	4(748)	5(928)	/	/	94(14043)
Publicité	3(858)	2(399)	/	/	56(8672)
Recette	1(173)	/	/	/	7(798)
Texte	283(76165)	156(63014)	175(89911)	48(34084)	1438(424009)
Total	321(86302)	198(71701)	184(92327)	49(34482)	2042(533192)

Tableau 4.2 – Nombre de textes par genre et par niveau, ainsi que le nombre de tokens (entre parenthèses).

On observe quelques tendances intéressantes, bien que peu surprenantes, sur les genres représentés dans les manuels. D'une part, plus le niveau des manuels augmente et plus la proportion de « Texte » augmente. Il peut s'agit d'extraits de romans, d'articles de journaux, de textes informatifs, etc. On voit donc que les débutants sont confrontés à une grande diversité de type de textes, laquelle diminue peu à peu, puis de manière drastique à la fin de B2. L'évolution la plus marquante concerne les dialogues : très présents au début de l'apprentissage (42,5 % des textes), ils sont nettement moins employés pour les apprenants intermédiaires et disparaissent aux niveaux C1 et C2. Cela semble confirmer l'hypothèse de Henry (1975) selon laquelle les dialogues constituent généralement des documents plus aisés.

	A1	A1+	A2	A2+	B1
Chanson	/	$0,8\%(0,6\%)$	$0,4\%(1,3\%)$	/	/
Dialogue	$41,3\%(39,7\%)$	$28,2\%(35,8\%)$	$17,6\%(22,6\%)$	$12,7\%(16\%)$	$6,7\%(6\%)$
E-mail	$5,1\%(5\%)$	$2,4\%(2,9\%)$	$1,7\%(1,1\%)$	$0,8\%(0,4\%)$	$2,2\%(1,1\%)$
Lettre	$6,1\%(3,9\%)$	$4,7\%(3,4\%)$	$3\%(2,1\%)$	$4,5\%(2,5\%)$	$7,2\%(9,1\%)$
Poème	/	$2,7\%(2,2\%)$	$2,6\%(2,3\%)$	$1,6\%(1\%)$	$0,6\%(0,4\%)$
Phrases	$13,8\%(14\%)$	$11,4\%(9,8\%)$	$5,6\%(6,2\%)$	$3,3\%(2,8\%)$	$2,2\%(1,3\%)$
Publicité	$3,1\%(3,4\%)$	$5,5\%(4,7\%)$	$6,4\%(2,8\%)$	$2,4\%(1,5\%)$	$2,8\%(3,1\%)$
Recette	$1\%(0,9\%)$	/	$1,3\%(0,7\%)$	$0,4\%(0,2\%)$	/
Texte	$29,6\%(33\%)$	$44,3\%(40,6\%)$	$61,3\%(60,9\%)$	$74,3\%(75,7\%)$	$77,8\%(78,8\%)$
Total	100%	100%	100%	100%	100%

	B1+	B2	C1	C2	Total
Chanson	/	/	/	/	$0,14\%(0,14\%)$
Dialogue	$4,7\%(6,6\%)$	$2,5\%(2,4\%)$	/	/	$13,2\%(10,1\%)$
E-mail	/	$3\%(1,9\%)$	/	/	$1,91\%(1,1\%)$
Lettre	$2,5\%(2\%)$	$6,1\%(3,9\%)$	$4,3\%(2,3\%)$	$2\%(1,2\%)$	$4,75\%(3,7\%)$
Poème	$2,2\%(1,1\%)$	$6,1\%(2\%)$	$0,5\%(0,3\%)$	/	$1,91\%(1\%)$
Phrases	$1,2\%(0,9\%)$	$2,5\%(1,3\%)$	/	/	$4,6\%(2,6\%)$
Publicité	$0,9\%(1\%)$	$1\%(0,6\%)$	/	/	$2,47\%(1,6\%)$
Recette	$0,3\%(0,2\%)$	/	/	/	$0,34\%(0,15\%)$
Texte	$88,2\%(88,3\%)$	$78,7\%(87,9\%)$	$95,1\%(97,4\%)$	$98\%(98,8\%)$	$70,4\%(79,5\%)$
Total	100%	100%	100%	100%	100%

Tableau 4.3 – Proportion de chaque genre de textes par niveau. Le pourcentage entre parenthèses représente la proportion de ces mêmes genres par niveau en termes de nombre de tokens.

4.3.6 Conclusion

Dans cette section, nous avons décrit la méthodologie employée pour rassembler le corpus annoté qui servira à entraîner et à valider notre formule de lisibilité pour le FLE. Notre approche postule qu'un texte extrait d'un manuel donné devrait être du même niveau de difficulté que celui-ci. Dès lors, comme les manuels de FLE récents sont étalonnés en fonction de l'échelle du CECR, il a été possible de numériser sur cette base pas moins de 2042 textes issus de 28 manuels différents. Ces textes ont été stockés dans des fichiers textes individuels, organisés au sein d'une architecture de répertoires. Ils ont également été enrichis à l'aide de quelques métadonnées qui permettent de connaître la source et le niveau de chaque document.

La répartition des textes par niveaux et par genre a révélé une domination croissante du genre « Texte » par rapport à des écrits plus typés à mesure que le niveau de compétence des apprenants augmente. Cette caractéristique ne représente toutefois pas un problème, étant donné que notre formule ne repose guère sur le type de textes, à l'exception des variables de dialogue (*cf.* Chapitre 3.3.4), et qu'il n'est pas nécessaire d'apprendre des régularités à ce niveau sur notre corpus.

Le nombre réduit de textes classés comme C2 nous semble par contre potentiellement plus problématique. En effet, il arrive que certains algorithmes statistiques se comportent moins bien lorsque les probabilités a priori de chacune des classes comprises dans le modèle ne sont pas identiques. Un second sujet délicat est le taux d'homogénéité du corpus. Dans notre discussion du CECR et de son échelle, nous avons souligné l'absence de régulation en ce qui concerne l'adéquation des niveaux affichés par les manuels et leur contenu réel. Dès lors, puisque cette décision est laissée à l'arbitraire de chaque équipe de rédacteurs, il est possible que l'on observe une certaine hétérogénéité entre manuels d'un même niveau. Ces deux problématiques que nous venons de soulever sont envisagées en détail dans la dernière section de ce chapitre. Celle-ci vise à analyser plus précisément ces deux aspects à l'aide d'outils statistiques et à proposer des solutions adaptées au diagnostic qui aura été posé.

4.4 Deux problèmes

La section précédente, qui décrivait notre corpus, a mis au jour deux questions délicates au sein de celui-ci : une possible hétérogénéité de l'étalonnage entre manuels d'un même niveau et le manque de textes de niveau C2. Cette troisième section aborde ces deux problématiques de façon plus systématique, en recourant tant à des analyses qualitatives que quantitatives. Les observations tirées de ces analyses aboutissent à la constitution de huit corpus différents, qui correspondent à autant de manières de résoudre ces deux problèmes.

4.4.1 L'homogénéité

Introduction

Afin d'évaluer l'homogénéité du corpus, nous avons recouru à une méthodologie qui consiste à sélectionner un nombre réduit d'indices de lisibilité représentatifs et à en comparer les moyennes obtenues pour chaque manuel. En ce qui concerne les indices, nous avons retenu deux variables classiques et relativement robustes : le nombre moyen de lettres par mots (**NLM**) [12] et le nombre moyen de mots par phrase (**meanNWS**). Ils constituent respectivement des indices de la difficulté lexicale et syntaxique des textes. Par ailleurs, nous avons aussi considéré la densité des idées (**ConcDens**) (*cf.* Chapitre 3.3.3) , afin d'obtenir une estimation de la complexité sémantique des textes.

12. Lors d'une première analyse, nous avions utilisé **ML1** comme indice de la difficulté lexicale. En effet, celui-ci s'était révélé plus discriminant que **NLM** (François, 2009a). Cependant, cet indice pose des problèmes spécifiques pour certains types de textes (petites annonces, recette de cuisine, etc.) dont la difficulté est surestimée. Comme ces genres sont plus répandus à certains niveaux qu'à d'autres, cette faiblesse introduit un biais sévère dans l'analyse des manuels. C'est pourquoi nous avons finalement préféré employer une variable plus robuste.

Dans un second temps, nous avons calculé les moyennes de chaque manuel pour ces trois variables, avant de les comparer. Ces comparaisons ont été opérées selon trois modes. Le premier repose sur une étude qualitative de plusieurs tableaux dans lesquels certaines irrégularités au niveau des moyennes sont mises en relief. Les deux tests subséquents sont de nature quantitative et visent à décider si les irrégularités observées sont suffisamment significatives pour conclure à la non-homogénéité du corpus. Ils s'appuient respectivement sur une analyse de la variance ANOVA, qui envisage chaque prédicteur individuellement et une analyse de la variance multivariée MANOVA, dans laquelle les effets des différents prédicteurs sont combinés.

La difficulté des manuels

La première étape de notre analyse a consisté à comparer les moyennes de chaque manuel à l'aide des trois variables retenues : **NLM**, **meanNWS** et **ConcDens**. Cette procédure est valide pour autant que ces trois indices soient suffisamment corrélés avec la difficulté des manuels, ce qui est le cas au moins pour **NLM** et **meanNWS**, comme nous l'avons montré dans François (2009a). Dès lors, (1) les moyennes de chaque indice pour chaque niveau devraient augmenter en fonction du niveau de difficulté des manuels. On peut alors postuler que si le corpus est homogène, (2) les moyennes de chaque manuel d'un même niveau seront assez proches de la moyenne globale du niveau et (3) au sein d'une même série, la moyenne de chaque manuel d'un niveau donné sera supérieure à l'ensemble des moyennes des niveaux inférieurs.

Afin d'explorer ces deux hypothèses, nous avons implémenté en *Python* une procédure automatisée qui applique chacune des trois fonctions de paramétrisation décrites dans le Chapitre 3 à l'ensemble des textes d'un niveau, tout en conservant la source de ces textes. Le fichier de résultats généré en sortie est ensuite analysé à l'aide d'un script *R*, afin d'obtenir diverses statistiques. Il faut noter que nous disposons de deux corpus différents : l'un comporte 6 échelons de difficulté (nous l'appelerons **Corpus6**), tandis que le second en a 9 (**Corpus9**). Par conséquent, nous rapportons ci-dessous les analyses pour chaque indice sur ces deux corpus.

Le corpus à 6 classes : Pour ce premier corpus, on observe plusieurs faits intéressants, rapportés dans les Tableaux 4.4, 4.5 et 4.6. Tout d'abord, l'hypothèse (1) semble se confirmer pour **meanNWS**, ainsi que pour **NLM** (à l'exception du niveau B1). Par contre, elle apparaît moins adéquate pour **ConcDens**, même si une évolution croissante est tout de même perceptible. Cela laisse penser que ce dernier facteur est moins fiable comme indice de la difficulté, car moins corrélé à cette dernière.

Afin de vérifier cette première observation, nous avons sélectionné 50 textes par niveau, en appliquant un échantillonnage stratifié au sein de chacun d'eux. Il s'agit d'une procédure qui sélectionne aléatoirement des observations tout en respectant leur distribution dans les différentes classes possibles, c'est-à-dire ici les différents manuels. Nous avons ensuite calculé les corrélations de Pearson (*cf.* Chapitre 5 pour les détails de la procédure) entre la difficulté de ces textes et nos trois variables explicatives. Celles-ci révèlent une nette supériorité de **meanNWS** ($r = 0,62$) et **NLM** ($r = 0,52$) sur **ConcDens** ($r = 0,37$), ce qui confirme nos craintes. C'est pourquoi nous ne traiterons guère de ce dernier facteur dans la suite de notre discussion.

	A1	A2	B1	B2	C1	C2
Activités	/	/	4,56	4,70	**4,61**	4,81
Alter Ego	4,37	4,42 (1) 4,48 (2)	4,60 (2) 4,61 (3)	/	/	/
Comp. écrite	/	**4,67**	4,67	/	/	/
Connexions	4,21	/	/	/	/	/
Conn. : prep. DELF	4,30	4,51	/	/	/	/
Delf/Dalf	/	/	/	4,64	4,80	4,88
Festival	4,41	/	4,63	/	/	/
Ici	4,40	4,68	/	/	/	/
Panorama	4,37	4,63 (1) **4,57 (2)**	4,69 (2) 4,68 (3)	**4,53**	/	/
Rond-point	**4,62**	4,61	**4,50**	**4,50**	/	/
Réussir Dalf	/	/	/	/	5	**4,97**
Taxi !	**3,92**	4,41	4,70	/	/	/
Tout va bien !	4,27	**4,25** (1) 4,41 (2)	**4,78 (2)** **4,69** (3)	/	/	/
Total	4,32	4,52	4,64	**4,58**	4,78	4,91

Tableau 4.4 – Nombre moyen de lettres par mots pour chaque manuel. Nombre moyen de mots par phrase pour chaque manuel. Les ouvrages présentant un problème de cohérence sont indiqués en gras. Les chiffres entre parenthèses réfèrent au numéro du manuel au sein de sa série. Comme on peut le noter, certains niveaux contiennent des textes issus de deux ouvrages différents de la même série.

En ce qui concerne la seconde hypothèse, elle apparaît valide dans la majorité des cas – du moins pour **meanNWS** et **NLM** –, même si l'on rencontre certains lieux d'achoppement représentés en gras dans les trois tableaux. Soulignons, au niveau lexical, les résultats élevés obtenus par *Rond Point* A1, *Comp. écrite* A2 et *Tout va bien !* B1 (2), et à l'inverse, les scores étonnamment faibles de *Taxi !* A1, *Tout va bien !* A2 (1), *Activités* C1, et même *Rond Point* B1 et B2. Au niveau syntaxique, on trouve surtout des ouvrages avec des phrases plus longues que la moyenne, tels que *Comp. écrite* A2, *Festival* B1, les exercices de préparation au DELF de *Connexion* et *Tout va bien !* A2 (1). Ce dernier ouvrage est intéressant parce qu'il mêle des phrases longues et des vocables simples. Il montre que les divergences observées par rapport aux moyennes peuvent certes s'expliquer par un problème de classification,

	A1	A2	B1	B2	C1	C2
Activités	/	/	18,2	19,6	**18,3**	21,9
Alter Ego	8,4	11,5 (1) 13,78 (2)	14,9 (2) 16 (3)	/	/	/
Comp. écrite	/	**17,1**	18,1	/	/	/
Connexions	10,1	/	/	/	/	/
Conn. : prep. DELF	12,9	**19,5**	/	/	/	/
Delf/Dalf	/	/	/	17	19,1	20,9
Festival	7,8	/	**19,9**	/	/	/
Ici	10,4	13,5	/	/	/	/
Panorama	8,6	10,6 (1) 12,4 (2)	**13,5** (2) 16,3 (3)	**16,5**	/	/
Rond-point	11,8	15,2	14,8	19,7	/	/
Réussir Dalf	/	/	/	/	21,4	21,7
Taxi !	7,6	15,2	16	/	/	/
Tout va bien !	9,9	**19,4** (1) **13,9** (2)	17,5 (2) 18,3 (3)	/	/	/
Total	9,1	14,54	16,85	18,6	19,36	21,43

Tableau 4.5 – Nombre moyen de mots par phrase pour chaque manuel. Les ouvrages présentant un problème de cohérence sont indiqués en gras.

mais aussi simplement parce que la progression au sein des différentes dimensions n'est pas conforme à la norme, sans que les textes ne soient globalement trop compliqués.

Quant à la progression au sein de chaque série, elle pose également certains problèmes, souvent au niveau de ces mêmes manuels qui s'éloignent de la moyenne de leur niveau. Ainsi, le pire des cas est *Rond Point*, dont la difficulté lexicale diminue entre A1 et B2. Ce profil « plat » se retrouve également pour *Comp. écrite*, aussi bien au niveau de **NLM**, **meanNWS** que **ConcDens**. Enfin, on note quelques endroits où une série qui apparaît assez cohérente dans l'ensemble connaît un problème localisé. C'est le cas de *Tout va bien !* A2 (1) dont nous avons déjà parlé, mais aussi de *Panorama* B2.

Il ressort que certaines méthodes sont plus problématiques que d'autres, en particulier *Comp. écrite*, et surtout, *Rond Point*. À notre sens, ce résultat s'explique par certaines caractéristiques de cette dernière : la progression y est rapide (elle vise sans doute des faux débutants) et opère selon une spirale où l'apprenant est rapidement mis en contact avec des formes plus complexes, qui ne sont cependant pas approfondies. Dès lors, les premiers textes sont fort complexes, mais par la suite, ce sont surtout les tâches associées qui deviennent difficiles et la complexité des textes semble moins évoluer que dans d'autres ouvrages. Voyons maintenant ce qu'il en est du corpus à 9 classes.

Le corpus à 9 classes : On rencontre assez naturellement davantage de problèmes dans le corpus à 9 classes, comme le révèlent les Tableaux 4.7 et 4.8 [13]. Tout d'abord, on retrouve

13. Nous n'avons pas présenté les valeurs pour la densité des idées pour le Corpus9, étant donné que cet indice s'est révélé peu intéressant par rapport à notre objectif, ainsi qu'il a été expliqué ci-dessus.

	A1	A2	B1	B2	C1	C2
Activités	/	/	0,464	0,465	0,473	**0,454**
Alter Ego	0,437	**0,476** (1) 0,458 (2)	**0,474** (2) 0,457 (3)	/	/	/
Comp. écrite	/	0,462	0,463	/	/	/
Connexions	0,423	/	/	/	/	/
Conn. : prep. DELF	**0,456**	**0,48**	/	/	/	/
Delf/Dalf	/	/	/	0,471	**0,48**	0,473
Festival	0,42	/	0,461	/	/	/
Ici	0,439	0,46	/	/	/	/
Panorama	0,417	0,447 (1) **0,432** (2)	**0,431** (2) 0,452 (3)	**0,446**	/	/
Rond-point	**0,457**	**0,443**	0,463	**0,452**	/	/
Réussir Dalf	/	/	/	/	0,472	0,479
Taxi !	0,426	0,458	0,466	/	/	/
Tout va bien !	**0,461**	0,45 (1) **0,467** (2)	**0,452** (2) **0,454** (3)	/	/	/
Total	0,43	0,457	0,459	**0,457**	0,475	**0.472**

Tableau 4.6 – Nombre moyen d'idées par texte pour chaque manuel. Les ouvrages présentant un problème de cohérence sont indiqués en gras.

logiquement les mêmes lieux d'achoppements que dans le Corpus6. Pour **NLM**, *Rond Point* A1 et A1+, *Comp. écrite* A2 et A2+ et *Tout va bien !* B1 (2) apparaissent à nouveau nettement au-dessus de la moyenne de leur niveau, tandis que *Taxi !* A1 et A2, *Tout va bien !* A2 (1), *Activités* C1 et *Rond Point* B1, B1+ et B2 semblent trop faciles. Cependant, le découpage des niveaux inférieurs a mis au jour d'autres lieux problématiques, tels que *Ici* A1, *Festival A1* ou *Tout va bien !* B1 (3).

Quant à **meanNWS**, on observe que *Conn. prep. DELF* A2, *Comp. écrite* A2, *Festival* B1 et *Tout va bien !* A2 (1) comportent des phrases plus longues que la moyenne, comme c'était le cas pour le Corpus6. Par ailleurs, comme pour **NLM**, des problèmes de découpage apparaissent. Un manuel dont la moyenne était dans la norme pour le Corpus6 devient problématique, une fois qu'il a été subdivisé en deux sous-niveaux, l'un dont les phrases sont étonnamment courtes et l'autre, trop longues. C'est le cas d'*Activités* B1 B1+, *Connexions* A1 A1+, *Panorama* A2 A2+ et *Tout va bien !* B1 B1+.

Enfin, en ce qui concerne l'hypothèse (3), l'aspect plat ou en dents de scie de *Rond Point* et de *Comp. écrite* est conservé, tandis que la classification des différents manuels au sein de *Tout va bien !* apparaît plus hétérogène que pour le Corpus6. Malgré tout, la majorité des autres séries présente un profil ascendant, aussi bien en termes lexicaux que syntaxiques.

En conclusion, on ne peut nier que l'homogénéité de notre corpus soit sujette à caution. Les profils « plats » de *Rond Point* et de *Compréhension écrite* sont particulièrement préoccupants. Cependant, dans l'ensemble, la plupart des séries respectent le profil ascendant que nous attendions, ou ne s'en écartent qu'à un moment particulier. De plus, il ne faut pas

	A1	A1+	A2	A2+	B1	B1+	B2	C1	C2
Activités	/	/	/	/	4,59	**4,52**	**4,7**	**4,61**	4,81
Alter Ego	**4,43**	4,33	4,42 (1) 4,48 (2)	4,47	4,6 (2) 4,6 (3)	4,62	/	/	/
Comp. écrite	/	/	**4,72**	**4,65**	**4,61**	4,72	/	/	/
Connexions	4,27	**4,14**		/	/	/	/	/	/
Conn. : prep. DELF	/	4,31	/	4,51	/	/	/	/	/
Delf/Dalf	/	/	/	/	/	/	4,64	4,8	4,88
Festival	**4,44**	4,38	/	/	4,62	4,64	/	/	
Ici	**4,51**	4,35	4,59	**4,81**	/	/	/	/	/
Panorama	4,18	**4,58**	**4,62** **4,71** (2)	4,52	4,69 (2) **4,62** (3)	4,7	4,53	/	/
Rond-point	**4,61**	**4,62**	**4,49**	**4,69**	**4,48**	**4,53**	4,5	/	/
Réussir Dalf	/	/	/	/	/	/	4,5	5	**4,97**
Taxi !	**3,88**	3,98	4,46	**4,36**	4,66	4,75	/	/	/
Tout va bien !	/	4,27	**4,25**	4,46	**4,78** (2) **4,58**	4,75	/	/	/
Total	4,29	4,33	4,52	4,54	4,62	4,67	**4,58**	4,78	4,91

Tableau 4.7 – Nombre moyen de lettres par mots pour chaque manuel. Les ouvrages présentant un problème de cohérence sont indiqués en gras. Les chiffres entre parenthèses réfèrent au numéro du manuel au sein de sa série. Comme on peut le noter, certains niveaux contiennent des textes issus de deux ouvrages différents de la même série.

	A1	A1+	A2	A2+	B1	B1+	B2	C1	C2
Activités	/	/	/	/	17,2	**19,2**	19,6	18,3	21,9
Alter Ego	6,6	10,2	11,5 (1) 13,8 (2)	13,8	14,9 (2) 15,8 (3)	16,1	/	/	/
Comp. écrite	/	/	**17,3**	**16,9**	**16,6**	**19,2**	/	/	/
Connexions	7,1	**13,9**	/	/	/	/	/	/	/
Conn. : prep. DELF	/	12,9	/	**19,5**	/	/	/	/	/
Delf/Dalf	/	/	/	/	/	/	17	19,1	20,9
Festival	6,6	9,3	/	/	**19,2**	**20,6**	/	/	
Ici	8,7	11,2	12,3	15,3	/	/	/	/	/
Panorama	6,9	10,5	10,6 (1) **9,3** (2)	13,6	13,5 (2) 15,7 (3)	16,5	16,5	/	/
Rond-point	6,6	12,6	14,9	15,2	**13,2**	16,6	19,7	/	/
Réussir Dalf	/	/	/	/	/	/	17	21,4	21,7
Taxi !	7,2	8,1	13,5	16,9	**15,4**	16,6	/	/	/
Tout va bien !	/	9,9	**19,4**	14,1 (2)	17,5 (2) **19,4**	17,6	/	/	/
Total	7	10,7	13,9	**18,1**	16,1	17,7	18,6	19,4	21,4

Tableau 4.8 – Nombre moyen de mots par phrase pour chaque manuel. Les ouvrages présentant un problème de cohérence sont indiqués en gras.

oublier que deux autres éléments viennent ajouter à la possible hétérogénéité des manuels :

1. Les variables que nous utilisons ne sont pas parfaitement corrélées avec la difficulté. Il est donc possible que certains points d'achoppement observés ne s'expliquent pas par une difficulté inappropriée, mais simplement par un score hors norme pour l'un des prédicteurs, lequel serait contrebalancé par d'autres aspects textuels que nous n'avons pas mesurés correctement (cohérence, densité des idées, etc.) ;

2. D'autre part, on observe davantage de problèmes dans le Corpus9, pour lequel nous avons dû prendre des décisions arbitraires en ce qui concerne le classement des textes. Il est possible que certaines d'entre elles soient inadéquates. Cependant, le cas d'*Alter Ego* montre que l'effet de ce facteur n'est pas systématique. Bien que la découpe exacte soit directement spécifiée dans les ouvrages pour les niveaux A1 à B1, le résultat de **NLM** pour A1 est quand même plus élevé que la moyenne du niveau.

Tests statistiques de l'homogénéité du corpus

Si l'analyse qualitative effectuée dans le point précédent a permis de dresser un tableau précis de la complexité de chaque manuel à l'aide d'indices lexicaux et syntaxiques, il est malaisé de trancher sur cette base quant à l'homogénéité du corpus. Les écarts observés sont-ils suffisants pour considérer qu'il y a un problème de cohérence dans l'annotation entre manuels, d'autant que la majorité des séries suivent tout de même une courbe ascendante ? Une façon plus systématique de répondre à cette question consiste à employer un test statistique testant le taux d'homogénéité au sein de chaque niveau de difficulté. Si celui-ci est insuffisant, le test rejettera cette hypothèse.

Étant donné que nos variables à expliquer sont quantitatives et que la variable explicative (les manuels) est nominale, le test adéquat pour évaluer une telle hypothèse est l'ANOVA (Howell, 2008, 305-352), pour autant que les distributions de **meanNWS** et **NLM** soient normales et leurs variances homoscédastes, c'est-à-dire égales. Dans le cas où les données s'éloignent fortement de la normalité, il est plutôt recommandé de recourir à un test non paramétrique tel que l'analyse de variance de Kruskal-Wallis (Howell, 2008, 680-681).

Afin de vérifier ces deux conditions, nous avons successivement appliqué les tests de normalité de Shapiro-Wilk (Shapiro et Wilk, 1965) et d'homoscédasticité de la variance de Levene (Brown et Forsythe, 1974) à chaque ensemble d'observations correspondant à un manuel [14]. Pour le Corpus6, l'hypothèse de la normalité n'a été rejetée que 27 fois sur 82 (dont 18 fois aux niveaux A1 et A2), tandis que le test de Levene n'a pas pu démentir l'homoscédasticité des variances pour 4 niveaux sur 12 [15]. Quant au Corpus9, seules les valeurs issues de 29 manuels sur 124 ont été considérées comme non normales [16], tandis que 3 tests de Levene sur 18 se sont révélés significatifs. Sachant que l'ANOVA tolère, dans une certaine mesure, que ses conditions soient violées, il nous a semblé plus justifié d'appliquer l'ANOVA à nos données qu'un test non paramétrique.

14. Pour l'ensemble des tests réalisés dans cette section, la valeur α du seuil de signification est $0,05$.

15. Nous rapportons ici de manière conjointe les tests pour **NLM** et **meanNWS**. En réalité, la normalité des distributions pour **NLM** est nettement plus nette que pour **meanNWS** (5 tests significatifs contre 22).

16. Là encore, on note une nette différence entre les distributions basées sur **NLM** dont seules 4 sur 62 sont non normales et celles basées sur **meanNWS**, où près de 25 obtiennent un rejet significatif de la normalité.

L'ANOVA est un test statistique couramment employé pour comparer plus de deux moyennes entre elles. Sans entrer dans les détails (cf. Howell, 2008, 305-352), elle considère que l'ensemble des observations X_{ij}, réparties dans les j traitements T (cad. les variables nominales dont les effets sur la variable continue sont étudiés ; ici, ce sont les manuels), peuvent être représentées à l'aide du modèle suivant :

$$X_{ij} = \mu + \tau_j + \varepsilon_i \tag{4.1}$$

Autrement dit, la variabilité d'une observation se décompose en deux types de variation :

1. la variabilité inter-traitement : l'écart τ_j entre la moyenne μ_j du traitement auquel elle appartient et μ, la moyenne globale ;
2. la variabilité intra-traitement : l'écart ε_i entre la valeur X_{ij} de l'observation et la moyenne de sa classe μ_j.

Dès lors, le test de l'ANOVA consiste à effectuer le ratio entre la variance inter-traitements totale sur la variance intra-traitements totale. Ce ratio, noté F, est distribué selon la loi de Fisher. Lorsqu'il atteint une valeur suffisamment élevée qui est fonction du seuil de signification α retenu, on peut rejeter l'hypothèse que l'ensemble des échantillons testés proviennent d'une même population. En effet, cela signifie que la variabilité inter-traitement est nettement plus importante que la variabilité intra-traitement et donc que l'influence de la variable explicative (ici, les manuels) sur la variable à expliquer (ici, **meanNWS** ou **NLM**) est significative. Dans ce cas, il faut en conclure que tous les manuels ne peuvent pas provenir d'une même population et qu'il y a donc des différences nettes entre la difficulté de manuels d'un même niveau.

Les résultats des différentes ANOVA pour chaque niveau du Corpus6 et du Corpus9 et pour nos deux indices de la lisibilité sont repris dans le Tableau 4.9 et sont plutôt préoccupants. Ainsi, pour **NLM**, on note que seuls B1 et C2 sont considérés comme homogènes dans le Corpus9 (et seul B1 pour le Corpus6). Quant à **meanNWS**, la situation est légèrement meilleure, puisqu'au moins 4 échelons (A1, B2, C1 et C2) seraient cohérents dans le Corpus9. Les différents problèmes soulignés lors de l'analyse qualitative se concrétiseraient donc bien. Le corpus que nous avons rassemblé ne serait donc pas parfaitement homogène et la plupart des niveaux comporteraient des manuels dont la difficulté varie sensiblement. Comme le révèlent les ANOVA, ce problème d'homogénéité apparaît surtout dans les niveaux médians : de A1+ à B1+.

Une question demeure cependant. L'ANOVA ne permet de comparer qu'une variable à la fois. Or, nous avons montré que certains manuels (par ex. : *Tout va bien ! A2*) s'éloignent de la norme par une dimension, mais qu'il la respecte par d'autres aspects. Par conséquent, n'est-il pas envisageable que considérer indépendamment chacun des deux indices augmente

	A1		A2		B1		B2	C1	C2
	A1	A1+	A2	A2+	B1	B1+	B2	C1	C2
NLM									
Corpus6	$***$		$***$		$0,02^*$		$0,02^*$	$***$	$0,39$
Corpus9	$***$	$***$	$***$	$0,01^*$	$0,45$	$0,01^*$			
meanNWS									
Corpus6	$**$		$***$		$***$		0.27	$0,11$	$0,82$
Corpus9	$0,65$	$0,06$	$0,04^*$	$0,02^*$	$***$	$***$			

Tableau 4.9 – P-valeur pour chacun des tests ANOVA effectués sur les deux corpus. Un résultat inférieur à $0,05$ *signifie que l'hypothèse d'homogénéité a été rejetée pour ce niveau en particulier. Les niveaux de significativité sont représentés comme suit :* $p < 0,001$ *:* $***$*;* $p < 0,01$ *:* $**$ *et* $p < 0,05$ *:* $*$*.*

l'impression d'hétérogénité du corpus ? Afin de vérifier cette dernière hypothèse, nous avons recouru à une version multivariée de l'ANOVA, la MANOVA, qui modélise l'influence de divers traitements sur plusieurs variables dépendantes en même temps. À nouveau, nous ne détaillerons pas cet outil, plus complexe encore que l'ANOVA (cf. Lewis-Beck, 1993, 340-368).

En réalité, les hypothèses posées par ce test sont encore plus strictes que dans le cas de l'ANOVA, puisqu'il faut non seulement que l'ensemble des moyennes de **NLM** soient similaires pour tous les manuels d'un niveau donné, mais que ce soit également le cas des moyennes pour **meanNWS**. En conséquent, pour les Corpus9 et Corpus6, seul le niveau C2 ne voit pas l'hypothèse nulle rejetée. Ce test ne nous apporte donc pas l'information voulue. Il ne permet pas de déterminer si un manuel est homogène au niveau lexical OU syntaxique, mais bien s'il l'est à la fois pour l'un ET pour l'autre.

Conclusion

Aussi bien l'analyse qualitative que les tests statistiques ont mis au jour un problème d'homogénéité du corpus, pour autant que les indices utilisés soient bien représentatifs de la difficulté. Ces résultats apparaissent préoccupants, car il n'est pas possible d'apprendre correctement des régularités à l'aide d'algorithmes statistiques sur un corpus dont l'annotation est incohérente.

Cependant, nos analyses ont également révélé que les problèmes se situent surtout au niveau de certains manuels et que le reste du corpus dénote une certaine cohérence. Nous avons dès lors pris la décision de constituer deux nouveaux corpus, **Corpus6Cleaned** et **Corpus9Cleaned** desquels ont été supprimés les textes issus des deux séries les plus problématiques : *Rond Point* et *Compréhension écrite*. Cependant, il ne semble pas que cette suppression solutionne le problème, comme le montrent les résultats des ANOVA calculées

sur le **Corpus6Cleaned**[17] présentées dans le Tableau 4.10. On peut noter que l'homogénéité des niveaux B2 et B1 a nettement augmenté pour **NLM**, au détriment de celle B2 pour **meanNWS**. Par conséquent, les quatre corpus présentés ici ont été conservés pour la suite de notre étude.

	A1	A2	B1	B2	C1	C2
NLM						
Corpus6Cleaned	$***$	$***$	$0,09$	$0,04^{*}$	$***$	$0,39$
meanNWS						
Corpus6Cleaned	$**$	$**$	$***$	0.01^{*}	$0,11$	$0,82$

Tableau 4.10 – P-valeur pour chacun des tests ANOVA effectués sur le nouveau Corpus6Cleaned. Un résultat inférieur à $0,05$ *signifie que l'hypothèse d'homogénéité a été rejetée pour ce niveau en particulier. Les niveaux de significativité sont représentés comme suit :* $p < 0,001$ *:* $***$ *;* $p < 0,01$ *:* $**$ *et* $p < 0,05$ *:* $*$.

4.4.2 Le nombre de textes par classe

La seconde particularité de notre corpus susceptible d'engendrer des problèmes est le faible nombre de textes annotés C2, qui s'explique par le peu de manuels disponibles pour ce niveau. Or, cela a été soulevé à la Section 4.3.1, l'un de nos objectifs en constituant ce corpus à partir de manuels de FLE est de disposer de suffisamment de textes pour pouvoir appliquer des techniques d'apprentissage automatisé. Sans vouloir entériner à tout prix l'adage « more data is better data », il nous est apparu que cette limitation pouvait nuire à l'entraînement de nos modèles.

Les probabilités a priori

Pour bien comprendre en quoi le manque de textes C2 peut représenter un obstacle, il nous faut revenir sur un aspect particulier de l'échantillonnage utilisé : les probabilités *a priori* des différents niveaux du corpus. Il s'agit de la probabilité qu'un texte rencontré par un lecteur provienne d'un niveau de difficulté donné plutôt qu'un autre. Ces probabilités peuvent être équiprobables ou non, mais, en général, certaines catégories d'observations sont plus fréquentes que d'autres au sein de la population. Ainsi, il est ardu de trouver des textes A1 ou A2 sur le web, tandis que les textes C1 et C2 y sont majoritaires.

Cependant, dans le contexte des manuels de FLE, les probabilités *a priori* sont déterminées par la disponibilité des manuels, ce qui ne constitue pas une donnée très naturelle. C'est pourquoi il nous apparaît que la question des probabilités *a priori* doit être plutôt guidée

17. Les résultats pour l'autre corpus étant fort similaires, ils ne sont pas présentés.

par des visées d'efficacité du modèle que par le respect d'un état de fait. Par conséquent, se pose la question suivante : est-il préférable de reconstituer un corpus dont les probabilités *a priori* de chaque niveau sont similaires ou de respecter les proportions obtenues au terme du processus de collecte décrit précédemment ?

Sur cette question, Hosmer et Lemeshow (1989, 147) déclarent : « Classification is sensitive to the relative sizes of the two component groups and will always favor classification into the larger group ». Ils traitent d'un modèle en particulier : la régression logistique, qui est décrite dans le Chapitre 5. Néanmoins, cette caractéristique est également attribuée à d'autres modèles par Tan *et al.* (2005, 294-304).

C'est également le constat que nous avions dressé lors d'une expérience préalable (François, 2009a). Celle-ci comparait les performances d'un modèle de régression logistique ordinale (cf. Agresti, 2002, 274-282) sur deux corpus différents : le premier comportait les premiers textes que nous avions alors rassemblés (A1 : 44 ; A2 : 45 ; B1 : 107 ; B2 : 50 ; C1 : 78 et C2 : 19), tandis que le second avait été obtenu par échantillonnage aléatoire simple de l'ensemble du corpus de manière à conserver environ 50 textes par niveau. Nous avions entraîné une formule de lisibilité classique [18] à partir des deux corpus et avions noté des performances supérieures pour le modèle entraîné sur le second corpus.

Cette étude présente toutefois quelques biais méthodologiques. Tout d'abord, le nombre de textes n'est pas équivalent dans les deux corpus (340 contre 300). Même si la différence est minime, il n'est pas exclu qu'un nombre supérieur de données influence l'efficacité des modèles (*cf.* Chapitre 5 pour une discussion de ce facteur). Ensuite, les textes de chaque niveau du premier corpus proviennent d'un seul manuel, alors que le second corpus a été échantillonné à partir de plusieurs ouvrages par niveau. Dès lors, sa représentativité apparaît sujette à caution. Troisièmement, soulignons que si les probabilités *a priori* du premier corpus sont bien non-équiprobables, elles ne respectent pas pour autant la distribution de notre corpus actuel. Enfin, l'évaluation des performances elle-même présente un écueil, quand bien même elles ont été obtenues via un processus de validation croisée à dix échantillons (*cf.* Chapitre 5). En effet, comme la proportion des textes dans les corpus de tests correspond à celles des corpus d'entraînement, on ne peut comparer l'efficacité des deux modèles sur une tâche commune.

Dès lors, nous avons répliqué cette expérience à l'aide d'une méthodologie qui évite ces écueils. Tout d'abord, nous avons veillé à ce que nos deux jeux de données comportent le même nombre de textes, à savoir 300. Pour le premier corpus (**Corp1**), nous avons implémenté un algorithme d'échantillonnage qui sélectionne aléatoirement, dans le Corpus6, un

18. Elle n'utilise que les deux prédicteurs classiques déjà employés dans l'analyse de l'homogénéité du corpus : NLM et meanNWS.

nombre donné de textes en respectant les probabilités *a priori* du corpus global [19]. Quant au second (**Corp2**), il a été constitué à l'aide d'un échantillonnage aléatoire simple durant lequel 50 textes ont été retenus par niveau. Dans les deux cas, la technique d'échantillonnage est supposée assurer une représentation équilibrée des différents manuels et ainsi éviter les second et troisième biais décrits précédemment. Enfin, nous avons constitué un corpus de test de 120 observations (20 textes par niveau) par échantillonnage aléatoire simple de notre corpus entier [20].

Sur la base de ces deux jeux d'observations, nous avons développé deux formules de lisibilité « baseline » (avec **NLM** et **meanNWS** comme variables) et comparé leurs performances en termes de proportion de la variance expliquée (R^2), à partir du corpus de test. Étonnamment, les résultats de notre nouvelle expérience ne vont pas dans le sens de la précédente. Comme le rapporte le Tableau 4.11, le R^2 du modèle calculé sur le corpus aux probabilités *a priori* non équiprobables est légèrement supérieur à celui de l'autre modèle, qu'il s'agisse de la valeur calculée sur le corpus d'entraînement, le corpus de test ou via une procédure bootstrap .632 (cf. Tufféry, 2007, 369-370) qui répète l'expérience sur 100 jeux de données légèrement différents, obtenus par rééchantillonnages du corpus d'entraînement.

	Corpus d'entraînement	Bootstrap .632	Corpus de test
Corpus6Equi	$0,40$	$0,39$	$0,41$
Corpus6Apriori	$0,43$	$0,42$	$0,43$

Tableau 4.11 – Pour chacun des deux modèles, sont présentés le R^2 estimé sur le corpus d'entraînement, à partir d'une procédure de bootstrap .632 et sur notre échantillon de test.

Cependant, cette apparente supériorité doit être nuancée, une fois que l'on y regarde de plus près. Déjà lors de l'expérience précédente, nous avions noté que le nombre élevé de textes B1 déformaient l'espace de régression. Dès lors, 55% des prédictions étaient rangées dans la classe B1 (dont 72% à tort), alors que les résultats du second modèle présentaient une répartition plus hétérogène des prédictions entre les niveaux. La seconde expérience reproduit ce motif de résultats, comme le montrent les deux matrices de confusion [21] (Tableaux 4.12 et 4.13).

La lecture de ces deux matrices révèle que le modèle *Corpus6Equi* a du mal à classifier correctement les observations des niveaux médians (B1, B2 et même C1) et favorise légère-

19. Dans ce cas, cela nous donne les fréquences suivantes : A1 : 66 ; A2 : 72 ; B1 ; 99 ; B2 : 29 ; C1 : 27 et C2 : 7.

20. Remarquons qu'en procédant ainsi, certains textes des corpus d'entraînement peuvent se retrouver dans le corpus de test. À l'exception du niveau C2, la probabilité d'un tel cas reste cependant assez faible et ne nous a pas semblé constituer un problème.

21. Une matrice de confusion est un outil utile pour évaluer un classifieur. Elle croise les prédictions effectuées par le modèle avec les résultats attendus et offre ainsi un panorama intéressant de la répartition des prédictions. Voir le Chapitre 5.4.3 pour une définition plus formelle.

	Prédictions					
Réelles	1	2	3	4	5	6
1	15	4	1	0	0	0
2	5	7	1	4	1	2
3	3	4	3	3	3	4
4	0	4	5	1	4	6
5	3	0	3	3	3	8
6	0	1	2	0	4	13
Total	26	20	15	11	15	33

Tableau 4.12 – Matrice de confusion pour le modèle entraîné sur Corpus6Equi

	Prédictions					
Réelles	1	2	3	4	5	6
1	14	5	1	0	0	0
2	6	8	5	0	0	1
3	1	5	13	0	1	0
4	1	1	16	0	1	1
5	0	3	13	0	4	0
6	0	0	11	0	9	0
Total	22	24	59	0	15	2

Tableau 4.13 – Matrice de confusion pour le modèle entraîné sur Corpus6Apriori

ment les niveaux externes (A1 et C2), lesquels sont également ceux où les prédictions sont les meilleures. Toutefois, le modèle *Corpus6Apriori* est bien plus tranché. Il effectue pas moins de 59 prédictions B1 (sur 120 textes), dont 72% sont erronées, comme c'était déjà le cas lors de l'expérience 1. Plus grave encore, ce modèle n'attribue jamais le niveau B2 à un texte, ce qui constitue un comportement intolérable pour une formule de lisibilité décente. Par conséquent, il s'avère qu'il est préférable d'utiliser un corpus dont les probabilités *a priori* sont équiprobables, ce qui réplique les conclusions de notre première expérience.

Le problème des textes C2

Revenons maintenant à notre question d'origine : pourquoi le faible nombre de textes annotés C2 peut-il représenter un problème ? La réponse est très simple : notre expérience précédente ayant démontré l'intérêt d'employer un même nombre de textes par niveau, la taille des corpus que nous pouvons dès lors employer est fonction de la quantité de textes dans la catégorie la moins peuplée, c'est-à-dire C2. Celle-ci ne comprend que 49 documents, ce qui limite la taille maximale de nos corpus à 300 textes (modèle à 6 niveaux) et 450 textes (modèle à 9 niveaux).

Cette limitation est d'autant plus regrettable que la quantité d'observations disponibles pour certains niveaux dépasse les 350 textes, sans qu'il nous soit possible de les utiliser, alors que les techniques d'apprentissage automatisé sont connues pour se comporter d'autant mieux que l'on dispose de plus de données. Par conséquent, nous avons pris la décision de diviser,

lorsque c'est possible, chacun des textes C2 en plusieurs fragments, afin de disposer de plus d'observations. Ce découpage a été opéré manuellement en respectant deux règles : (1) ne pas découper un texte au milieu d'un paragraphe, afin de maintenir une certaine cohérence thématique au sein de chaque fragment ; (2) conserver des fragments de taille suffisante, c'est-à-dire comptant au moins une centaine de mots. Par conséquent, il n'a pas été possible de découper l'ensemble des textes, certains étant trop courts pour ce faire.

Ce faisant, nous sommes tout de même passé de 49 à 108 observations, qui se répartissent de la manière suivante entre les trois manuels : *Activités* : 15 fichiers ; *Dalf C1/C2* : 51 et *Réussir le Dalf* : 42. Cette modification nous a permis de définir quatre corpus supplémentaires :
– **Corpus6Ext** : le Corpus6 où certains des textes C2 ont été divisés en 2 à 5 fragments ;
– **Corpus6ExtCleaned** : le Corpus6Cleaned où la même opération a été effectuée ;
– **Corpus9Ext** : idem pour le Corpus9 ;
– **Corpus9ExtCleaned** : idem pour le Corpus9Cleaned.
Dès lors, il nous sera possible de constituer des corpus d'entraînement composés de respectivement 660 ou 990 textes et de vérifier si cette augmentation de taille conduit à un gain de performances significatif par rapport à notre modèle précédent (François, 2009a).

4.4.3 Conclusion

Dans cette troisième section, nous avons abordé deux problèmes essentiels de notre corpus : (1) un manque d'homogénéité dans les annotations, qui risque de nuire à l'apprentissage automatique des règles de classification par les algorithmes statistiques ; (2) le fait qu'il y ait trop peu de textes C2, ce qui nous empêche d'utiliser l'ensemble de nos données. En effet, nous avons démontré l'importance de conserver des probabilités a priori similaires, du moins tant que l'on ignore la distribution réelle des textes disponibles pour les professeurs de FLE.

Nous avons répondu à ces deux difficultés en effectuant des modifications dans les deux corpus de base : Corpus6 et Corpus9. D'une part, les séries les plus problématiques (*Rond Point* et *Compréhension écrite*) ont été écartées, ce qui nous donne deux nouveaux corpus : Corpus6Cleaned et Corpus9Cleaned. D'autre part, les textes C2 disponibles ont été fragmentés afin d'augmenter le nombre d'observations disponibles pour nos futures analyses. Afin d'évaluer les différentes modifications apportées à l'aulne de nos modèles, nous avons conservé les quatre corpus précédents, tout en définissant quatre variantes où l'on trouve 108 fragments C2 à la place de 49. Il s'agit de Corpus6Ext, Corpus6ExtCleaned, Corpus9Ext et Corpus9ExtCleaned. Cela nous donne donc un total de 8 jeux de données différents.

4.5 Conclusion générale

Confronté à la nécessité de rassembler un corpus de textes annotés en fonction de leur difficulté à la lecture pour des apprenants de FLE, nous avons passé en revue l'ensemble des critères employés en lisibilité pour évaluer le taux de compréhension des lecteurs pour un texte donné. L'emploi de tests de compréhension ou de closure a été rejeté tant pour des raisons théoriques que pratiques. Le temps de lecture, pourtant digne d'intérêt, n'a pas pû être utilisé par manque des ressources nécessaires à sa mise en œuvre pour un large ensemble de sujets et de textes. Dès lors, nous avons finalement opté pour des jugements d'experts, lesquels sont notamment disponibles en grand nombre dans les manuels de FLE. Ce faisant, nous nous sommes éloigné de la vision classique en lisibilité, à l'instar de Dolch (1948) et de la majorité des chercheurs liés à la lisibilité computationnelle. En effet, nous ne déterminons plus la difficulté de matériaux par rapport à une population de référence dont les compétences sont testées via des tests, mais nous cherchons à classer de nouveaux textes conformément aux normes en vigueur dans les ouvrages pédagogiques disponibles pour l'enseignement de population d'intérêt.

Une fois ce critère retenu, nous avons réfléchi au choix d'une échelle de difficulté. C'est celle du CECR, désormais largement répandue dans l'enseignement du FLE, qui s'est naturellement imposée, d'autant que la majorité des manuels de FLE actuellement disponibles situent leurs contenus d'après celle-ci. Dès lors, nous avons pu rassembler des textes annotés par des experts en appliquant la règle suivante : le niveau d'un texte donné est le même que celui du manuel dont il est issu. Sur cette base et une série de critères de sélection, nous avons sélectionné et annoté 2042 textes issus de 28 manuels et de 2 compléments accessibles sur le web. Chacun d'eux a été sauvegardé au sein d'un fichier commençant par un en-tête qui spécifie quelques méta-informations d'intérêt.

Dans une troisième étape, nous avons analysé les caractéristiques du corpus rassemblé et avons noté deux problèmes principaux. D'une part, annoter les textes en se basant sur le niveau des manuels semble engendrer une certaine hétérogénéité. Celle-ci s'explique à la fois par la variété de jugements que les auteurs des différentes collections ont émis quant à ce qui fait la difficulté de textes, mais aussi par le fait que le placement d'un texte à un niveau donné ne dépend pas seulement de la complexité intrinsèque de celui-ci, mais est influencé par la tâche qui lui est associée. Il arrive donc que certains textes issus d'un manuel de difficulté X soient en fait plus simples ou plus complexes que le niveau indiqué sur la couverture, parce qu'ils sont associés à une tâche respectivement plus ardue ou plus aisée qui compense cet écart. D'autre part, nous avons aussi montré l'importance d'utiliser un corpus dont les probabilités *a priori* de différents niveaux sont équilibrées, sans quoi le classifieur qui en est dérivé présente un biais envers certains niveaux au détriment d'autres.

Nous avons présenté diverses stratégies pour faire face à ces deux défauts, ce qui a conduit à la constitution de huit corpus différents. Ceux-ci vont être utilisés (1) pour analyser les régularités, dont nous avons déjà parlé, qui sont susceptibles d'exister entre la difficulté des textes et les prédicteurs définis au Chapitre 3 et (2) pour comparer divers algorithmes statistiques afin de déterminer le plus adéquat à notre objectif : définir une formule de lisibilité pour le FLE. Ces deux aspects sont présentés dans la troisième et dernière partie de cette thèse.

Troisième partie

Expérimentations et résultats

CHAPITRE 5

RÉSULTATS

5.1 Introduction

Cette troisième et dernière partie de notre étude prolonge les acquis méthodologiques de la partie précédente. Cette dernière a dressé l'inventaire des indices présumés de la difficulté, parmi lesquels 406 prédicteurs ont été retenus pour implémentation. Elle a aussi passé en revue les diverses manières de rassembler un corpus de textes annotés en termes de difficulté, ce qui a permis de sélectionner celle qui est apparue comme la plus adéquate à notre problématique. Dès lors, 2 042 textes ont été extraits de manuels de FLE, donnant le jour à huit corpus légèrement différents.

La dernière étape de notre démarche, décrite dans ce chapitre, consiste à utiliser ces apports théoriques et ces données afin de développer une formule de lisibilité pour le FLE, tout en prenant aussi position par rapport à plusieurs problématiques soulevées dans la partie théorique de cette étude. Ce programme a été appliqué en deux phases, qui correspondent aux deux sections principales de ce chapitre.

Dans un premier temps, une série d'analyses bivariées ont étudié la capacité prédictive des 406 variables, considérées une à une dans leur relation avec la difficulté. Autrement dit, nous avons évalué dans quelle mesure chaque variable respectait les qualités d'un bon prédicteur, lesquelles ont été décrites au Chapitre 3.2.1. Cette première étape a permis de détecter les facteurs les plus associés avec le niveau des textes des huit corpus et, par conséquent, les plus susceptibles d'être utiles au sein de notre formule.

Cependant, lorsqu'on veut combiner plusieurs facteurs au sein d'un modèle prédictif, leur apport informatif et leur significativité varient en fonction des autres facteurs déjà présents dans le modèle. En effet, deux variables fortement intercorrélées apportent des informations redondantes au modèle. Dès lors, il a été nécessaire, dans la seconde partie de ce chapitre, de déterminer quelle combinaison de prédicteurs apporte l'information la plus variée possible.

Pour ce faire, divers algorithmes statistiques et techniques de sélection de variables ont été comparés, ce qui a débouché sur une nouvelle formule pour prédire la lisibilité du FLE, dont la présentation clôt ce chapitre.

5.2 Analyse de la capacité prédictive des variables explicatives

5.2.1 Introduction

Lorsqu'on cherche à combiner diverses variables explicatives pour former un modèle prédictif, la première étape consiste à bien étudier le jeu de données dont on dispose. Il existe dans ce but plusieurs méthodes ressortissant au domaine de la fouille de données (cf. Tufféry, 2007). Certaines d'entre elles sont relativement complexes, telle l'analyse factorielle ; d'autres sont beaucoup plus simples et offrent des outils de visualisation et de synthèse des données, parmi lesquels on compte les indices de corrélation. Ceux-ci permettent entre autres d'appréhender les relations que les différentes dimensions des observations entretiennent avec le phénomène à expliquer, ce qui constitue justement le point qui nous intéresse particulièrement dans cette première section. Celle-ci vise en effet à analyser la capacité prédictive des 406 facteurs identifiés dans le Chapitre 3.

Afin de réaliser cet objectif, nous allons tout d'abord discuter les outils statistiques adaptés à l'exploration des relations entre variables (*cf.* Section 5.2.2). Ensuite, la procédure de constitution des différents corpus utilisés pour calculer ces mesures sera détaillée à la Section 5.2.3), avant de passer à l'analyse des variables à proprement parler. Afin d'éviter de perdre le lecteur parmi la masse de résultats, nous avons organisé la discussion selon la typologie décrite dans le Chapitre 3. Ainsi, l'exposé commence par traiter des indices de la difficulté lexicale (*cf.* Section 5.2.4), puis des aspects syntaxiques (*cf.* Section 5.2.5), avant de s'étendre sur les deux dimensions de la complexité sémantique que sont le contenu et la cohérence (rassemblées à la Section 5.2.6). Nous terminons cette partie par l'analyse des variables spécifiques au processus de lecture en L2 (*cf.* Section 5.2.7).

5.2.2 Outils de mesure d'une bonne variable

Comme le chapitre 2 l'a montré, le principal moyen utilisé depuis les débuts de la lisibilité pour évaluer l'efficacité d'un prédicteur consiste à calculer son degré de corrélation avec la difficulté des textes. Dans la majorité des cas, le coefficient de corrélation utilisé est celui de Bravais-Pearson. Il permet de mesurer le degré auquel deux variables aléatoires X et Y varient ensemble et fournit dès lors une bonne indication du pouvoir prédictif de X (la variable indépendante) par rapport à Y (la variable à prédire). Il se calcule à l'aide de la

formule suivante (Howell, 2008, 244) :

$$r = \frac{s_{XY}}{s_X s_Y} \tag{5.1}$$

où s_X et s_Y sont les écart-types de X et Y estimés à partir d'un échantillon et s_{XY} est la covariance de X et Y, également estimée sur cet échantillon. Pour un échantillon à N observations, la version développée de cette formule donne :

$$r = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2}} \tag{5.2}$$

où $\overline{x}$ et $\overline{y}$ correspondent aux moyennes de X et Y, estimées sur l'échantillon.

Au-delà de sa simplicité, cet outil statistique a déjà largement fait ses preuves dans la littérature en lisibilité. Cependant, la validité de son application est tributaire du respect de deux conditions portant respectivement sur l'échelle de mesure de X et Y et sur la nature de la relation qui les unit. Celles-ci sont décrites en détail dans les deux points suivants, qui s'attachent à déterminer dans quelle mesure nos propres données les respectent.

L'échelle de mesure des variables

La première de ces conditions spécifie que le coefficient de corrélation de Pearson s'applique à des données quantitatives. Or, si l'ensemble de nos 406 prédicteurs sont des variables continues ou binaires, comment faut-il considérer la difficulté des textes ? Stevens (1946, 1968) a formalisé quatre échelles de mesure – nominale, ordinale, intervalle ou ratio –, définies en fonction des transformations mathématiques qu'elles autorisent sans que ne soit altérée la signification des statistiques ensuite calculées. Cette définition revient à mettre en rapport chaque type de données avec une série de tests statistiques autorisés. Dès lors, sélectionner l'échelle adaptée à la mesure de la difficulté apparaît comme un prérequis à la sélection d'une mesure de la corrélation, mais aussi d'un modèle statistique.

Qu'en est-il donc de la difficulté des textes ? D'un point de vue théorique, l'échelle ordinale constitue le candidat le plus évident dans notre cas. Il s'agit de l'option habituellement retenue par les institutions d'enseignement des langues qui organisent le parcours d'apprentissage en G classes successives. Cette option semble raisonnable. D'une part, on ne peut pas considérer les niveaux du CECR comme de simples étiquettes interchangeables (propriété d'une variable nominale). Il y a bien une notion de progression : les textes du niveau B1 sont plus complexes que ceux du niveau A2, mais moins que ceux du niveau C1. D'autre part, cette organisation en niveaux ne nous dit rien de la distance qui sépare chacun d'eux,

c'est-à-dire de la quantité d'efforts que l'apprenant lambda doit accomplir pour passer d'un niveau donné au niveau suivant. Il est raisonnable de penser que cet intervalle entre deux niveaux d'éducation n'est pas constant. Ces deux arguments vont donc dans le sens d'une variable ordinale.

Toutefois, il existe des raisons qui justifient le recours à l'échelle par intervalles (c'est-à-dire quantitative discrète). Comme l'a expliqué Agresti (2002, 277-278), une variable ordinale Y à G niveaux cache souvent une variable latente continue Y^* qui a été discrétisée à l'aide de $G+1$ points de coupures $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_G = \infty$ selon la relation suivante :

$$Y = g \qquad \text{si } \alpha_{g-1} < Y^* \leq \alpha_g \tag{5.3}$$

Or, cette remarque nous semble particulièrement bien s'appliquer à notre situation où l'existence d'une telle variable latente est étayée par les deux faits suivants. D'une part, l'écart entre les premiers et les derniers textes du niveau A1 est supérieur à l'écart qui existe entre ces derniers textes du niveau A1 et les premiers du niveau A2. Cette particularité s'explique aisément si l'on considère la difficulté d'un texte comme continue et les séparations entre niveaux comme des points de coupure, très peu de choses distinguant alors les textes situés près de ceux-ci. D'autre part, cette hypothèse d'une variable latente peut expliquer les différences observées entre manuels dans le chapitre 4 : il s'agirait d'un désaccord sur les valeurs optimales des points de coupure α_j.

On ne peut dès lors se décider sur l'échelle la plus adaptée à partir des seuls arguments théoriques. D'autant que ce questionnement pourrait bien s'avérer inapproprié. En effet, bien qu'elle ait été très utilisée en statistiques, la typologie de Stevens a fait l'objet de nombreuses critiques. Velleman et Wilkinson (1993, 69) ont ainsi défendu que le type d'échelle n'est pas une propriété intrinsèque d'un phénomène, mais dépend de la question avec laquelle le chercheur approche ce phénomène. En lisibilité, cette vue avec laquelle le chercheur aborde la difficulté d'un texte revient en réalité souvent à la nature du critère qu'il emploie. C'est ainsi que la majorité des formules traditionnelles, basées sur les résultats de tests de compréhension à la lecture, ont considéré la difficulté selon une échelle à intervalles. Plus récemment, les travaux qui appliquent des techniques de l'intelligence artificielle à la lisibilité (cf. Heilman *et al.*, 2008) ont plutôt envisagé la difficulté comme une variable ordinale, voire nominale. Il apparaît donc valable de considérer la difficulté soit selon une échelle ordinale, soit une échelle à intervalles.

Afin de satisfaire cette première condition, il a fallu considérer les 6 ou 9 échelons de notre échelle de difficulté comme une variable quantitative discrète à 6 ou 9 valeurs. Cette opération revient à postuler une distance égale entre chaque niveau, ce qui constitue tout de

même une hypothèse forte. Dès lors, nous avons également adopté l'autre option : traiter la difficulté selon une échelle ordinale et transformer alors chacun de 406 prédicteurs en une variable ordinale. Dans ce cas, il convient d'employer une autre mesure de corrélation, qui se base sur les rangs, mais n'émet aucune hypothèse sur l'écart entre chaque niveau de la variable. Ayant le choix entre le coefficient de Spearman (Spearman, 1904) et de Kendall (Howell, 2008, 295-296), nous avons opté pour le premier, car celui-ci se calcule simplement en appliquant la formule de Pearson (*cf.* Équation 5.2) aux données rangées, c'est-à-dire où les valeurs ont été remplacées par leur rang dans la distribution empirique. Cette seconde mesure de corrélation permettra de prendre en compte des relations monotoniques croissantes, mais qui ne sont pas nécessairement linéaires, telles que les relations logarithmique ou exponentielle.

Linéarité des variables

La seconde condition à remplir pour que le coefficient de corrélation de Pearson mesure adéquatement la relation entre deux variables est que cette relation soit linéaire. Les Figures 5.1 et 5.2, obtenues à partir de données simulées avec R (R Development Core Team, 2008), sont révélatrices du problème qui survient en cas de non-linéarité. Dans le cas linéaire (*cf.* Figure 5.1), on voit que la droite en rouge, représentant la relation entre X et Y modélisée par la corrélation, constitue une bonne estimation de la manière dont les deux facteurs co-varient. À l'inverse, dans le cas non linéaire représenté à la Figure 5.2, X et Y co-varient également fortement, mais suivant une courbe représentée en bleu et obtenue à l'aide d'un modèle de régression polynomiale locale et pondérée (cf. Tufféry, 2007, 299-303). On remarque que, dans ce cas, la corrélation de Pearson (figurée en rouge) n'est pas capable de mesurer correctement l'association des deux variables. Le résultat obtenu sous-estime en effet la force d'association entre les deux facteurs.

En lisibilité, la majorité des chercheurs ont fait l'hypothèse que leurs différents indices entretenaient des rapports linéaires entre eux, ce qui justifiait l'emploi de la corrélation de Pearson comme moyen d'évaluation. Quelques-uns (Bormuth, 1966; Dale et Chall, 1948a; McLaughlin, 1969; Ojemann, 1934) ont cependant signalé que ce présupposé était parfois incorrect et ont adapté leurs formules afin de tenir compte de la relation non linéaire entre certains de leurs prédicteurs et leur critère. Comme nous l'avons dit au Chapitre 2, c'est Bormuth (1966) qui a le mieux exploré cette question.

Ce chercheur a détecté que la linéarité entre le taux de compréhension mesuré via un test de closure et ses divers prédicteurs évoluait en fonction de l'unité linguistique considérée (Bormuth, 1966, 98-102). Ainsi, au niveau des mots, il remarque que ni le nombre de syllabes ou de lettres, ni la fréquence ne sont corrélés linéairement avec la compréhension. Au

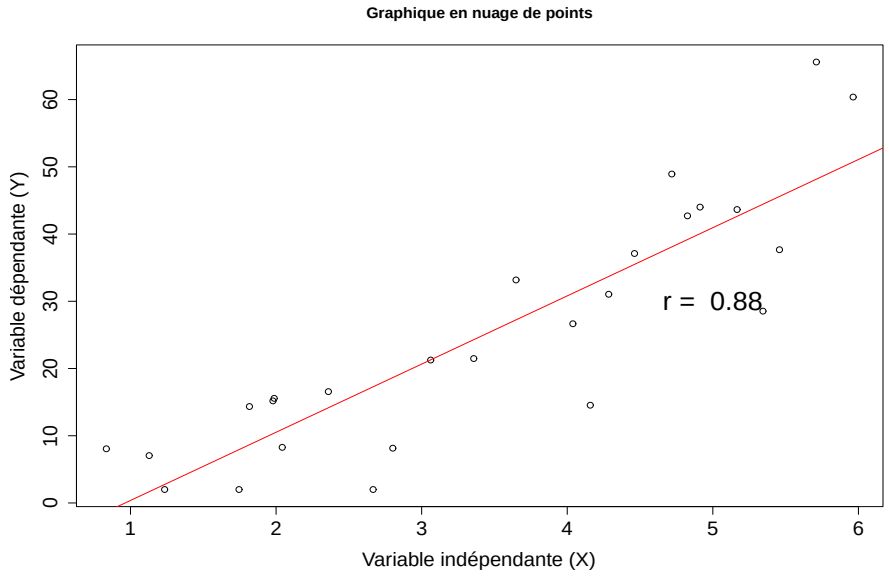


Figure 5.1 – Coefficient de corrélation entre deux variables en relation linéaire

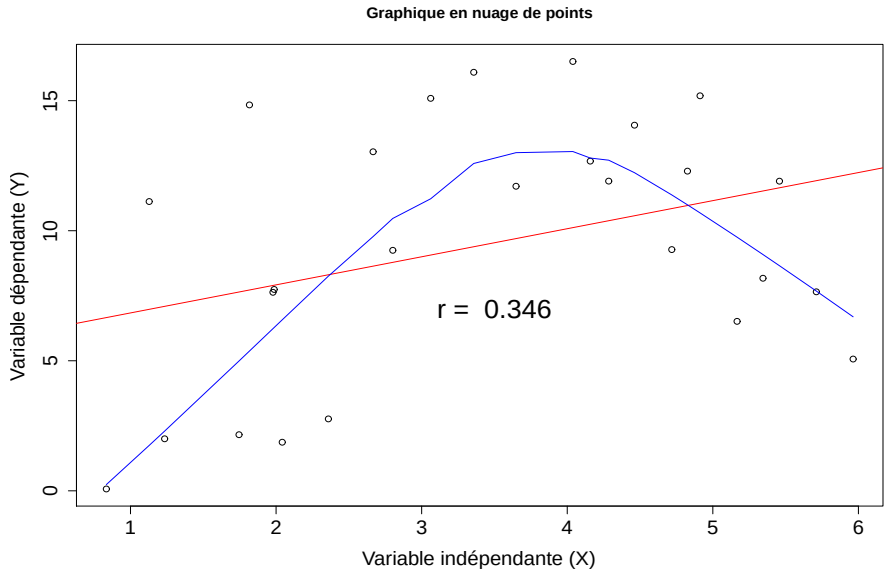


Figure 5.2 – Coefficient de corrélation entre deux variables en relation non linéaire

contraire, il observe une évolution curvilinéaire où la courbe est d'abord pentue, puis tend peu à peu vers un palier. Ce résultat fait écho à divers travaux sur la fréquence qui attestent, aussi bien dans le cadre de tâches de reconnaissance des mots (Balota et Chumbley, 1984; Just et Carpenter, 1980; Whaley, 1978) qu'en lisibilité (Stenner, 1996), du gain prédictif lorsqu'on emploie le logarithme des probabilités plutôt que la probabilité elle-même. Au niveau de la phrase, les relations tendent néanmoins à être davantage linéaires, à l'exception du nombre d'absents de la liste de 3 000 mots de Dale. Quant au niveau du texte, il voit encore se renforcer la linéarité des prédicteurs.

Pour Bormuth (1966, 84), cette tendance à la linéarité s'explique par le fait que, si les variables utilisées (par exemple, le nombre de lettres dans un mot) ne sont pas normales au niveau du mot, lorsqu'on passe à un niveau supérieur, l'indicateur obtenu est le résultat de la moyenne de plusieurs observations. Or, par le théorème central limite, une série de moyennes de variables non normales tend à adopter une distribution gaussienne et la relation entre deux variables gaussiennes est habituellement linéaire.

Afin de déterminer concrètement ce qu'il en est de nos différents prédicteurs, nous avons décidé de recourir au test F de linéarité décrit par Guilford (1965, 314). Celui-ci est basé sur la comparaison de deux mesures de corrélation : r^2 et η^2. La seconde, η^2, aussi appelée rapport de corrélation, est définie comme « le coefficient de corrélation associé à la régression curvilinéaire, c'est-à-dire la régression où la ligne la mieux ajustée n'est pas une ligne droite » (Howell, 2008, 336). Autrement dit, il exprime la proportion de variance d'Y qui pourrait être expliquée à l'aide d'une régression non linéaire adéquate. Il se calcule, dans le cadre d'une ANOVA, comme la différence entre la somme des carrées de la variabilité totale (SC_{total}) et la somme des carrés des erreurs résiduelles (SC_{erreur}), normalisée par SC_{total} (Howell, 2008, 337) :

$$\eta^2 = \frac{SC_{total} - SC_{erreur}}{SC_{total}} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{I}(Y_{ij} - \overline{Y})^2 - \sum_{j=1}^{J}\sum_{i=1}^{I}(Y_{ij} - \overline{Y}_j)^2}{\sum_{j=1}^{J}\sum_{i=1}^{I}(Y_{ij} - \overline{Y})^2} \tag{5.4}$$

où pour une variable Y à J traitements (ou valeurs), $\overline{Y}$ représente la moyenne générale de l'échantillon et $\overline{Y}_j$, la moyenne pour les observations appartenant au traitement j.

Sachant qu'η^2 est capable de prendre en compte l'association non linéaire de deux séries de données, sa valeur est généralement supérieure à celle de r^2. Guilford a dès lors défini un test de linéarité qui postule qu'une trop large différence entre ces deux mesures est indicatif d'une relation non-linéaire. Il se calcule de la façon suivante (Guilford, 1965, 314) :

$$F(k-2, N-k) = \frac{(r^2 - \eta^2)(N-J)}{(1-\eta^2)(J-2)} \tag{5.5}$$

où N est le nombre d'observations au total. On remarque que ce test produit une valeur distribuée selon la loi de Fisher, qui s'interpète en fonction des degrés de liberté suivants : $J-2$ et $N-J$. Suivant Bormuth (1966) sur ce point, nous avons décidé de reprendre ce test dans le but de vérifier la relation qu'entretient chacun de nos prédicteurs avec la difficulté des textes de notre corpus.

La normalité des variables

Une dernière caractéristique de nos prédicteurs qui mérite d'être évaluée dans cette section est leur normalité, et ce, pour deux raisons. D'une part, d'après Bormuth (1966), celle-ci influencerait le niveau de linéarité entre X et Y, puisque des variables gaussiennes tendent à entretenir une relation linéaire. D'autre part, la normalité des données constitue une condition requise par certains modèles statistiques, en particulier la régression linéaire.

Il existe plusieurs tests pour évaluer la normalité d'un échantillon : Kolmogorov-Smirnov, Lilliefors, Anderson-Darling, Shapiro-Wilk, etc. (cf. Rakotomalala, 2008). Parmi ceux-ci, le test de Shapiro-Wilk, déjà utilisé dans le chapitre 4, est reconnu comme étant l'un des plus puissants, c'est-à-dire qu'il n'hésite donc pas à rejeter l'hypothèse de normalité, à la différence d'autres tests plus conservateurs. Il utilise la statistique *W*, calculée de la façon suivante (Rakotomalala, 2008, 14) :

$$W = \frac{[\sum_{i=1}^{[\frac{N}{2}]} a_i(x_{(n-i+1)} - x_{(i)})]^2}{\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{5.6}$$

où $x_{(i)}$ correspond à l'observation i de la série des données rangées ; $[\frac{N}{2}]$ est la partie entière du rapport $\frac{N}{2}$ et a_i « sont des constantes générées à partir de la moyenne et de la matrice de variance co-variance des quantiles d'un échantillon de taille n suivant la loi normale » (Rakotomalala, 2008, 14). Plus la valeur de W est élevée, plus la distribution évaluée par ce test est proche de la normalité. Désirant obtenir cette information pour chacun de nos prédicteurs, nous les avons également soumis à ce test.

Au final, le dispositif exploratoire mis en place pour évaluer la capacité prédictive des 406 variables comprend quatre statistiques : (1) la corrélation de Pearson (r), qui évalue le niveau d'association linéaire entre un prédicteur X et la difficulté ; (2) la corrélation de Spearman (ρ), apte à mesurer une relation qui croît de façon monotone ; (3) la statistique F, visant à détecter des relations significativement non linéaires entre X et Y et (4) le W de Shapiro-Wilk, qui mesure la normalité d'une variable. Avant de détailler, pour chaque famille de variables, les résultats de ces quatre mesures, il reste à décrire les corpus de tests ayant servi à les calculer.

5.2.3 Constitution des corpus de tests

Une fois sélectionnés les outils statistiques de l'analyse, reste à les appliquer aux données. Le Chapitre 4 ayant mis en évidence l'importance d'entraîner un modèle statistique sur un corpus où la probabilité *a priori* de chaque classe est équilibrée, il est donc logique d'évaluer

l'efficacité des prédicteurs sur des corpus de même nature que ceux qui serviront à développer notre formule. La technique mise en œuvre pour constituer chacun d'eux opère en deux étapes. Dans un premier temps, une procédure de rééchantillonnage sélectionne aléatoirement un nombre équivalent de textes par niveau. Ensuite, les jeux de données obtenus sont nettoyés de leurs données aberrantes. Les détails de ces deux phases sont dévoilés à la suite de cette introduction. Par ailleurs, nous comparons dans un troisième point les huit corpus à notre disposition afin de retenir les deux meilleurs.

Rééchantillonnage des corpus de tests

Rééchantillonner revient à sélectionner aléatoirement un nombre réduit d'observations au sein d'un corpus. Dans notre cas, l'opération a été réalisée en deux étapes. Tout d'abord, l'ensemble des documents a été paramétrisé de sorte que chaque texte soit représenté sous la forme d'un vecteur de 406 variables explicatives. Ces vecteurs ont été sauvegardés dans un fichier unique dans lequel un script Python a permis, dans un second temps, de sélectionner aléatoirement un nombre M de textes par niveau, générant ainsi un échantillon de test dont les probabilités *a priori* sont identiques. Étant donné les limites imposées par le faible nombre de textes C2, M a été fixé à 50 pour les 4 versions non étendues du corpus (Corpus6, Corpus6Cleaned, Corpus9 et Corpus9Cleaned) et à 108 pour les quatre autres (Corpus6Ext, Corpus6ExtCleaned, Corpus9Ext et Corpus9ExtCleaned).

Le traitement des données aberrantes

Sur la base de ces huit corpus, il a été possible d'obtenir huit échantillons de tests différents qui ont ensuite été nettoyés de leurs données aberrantes (ou « outliers »). En effet, rappelons que le corpus rassemble des textes de natures très diverses (*cf.* Chapitre 4.3.5). Or, des expérimentations préalables ont révélé que certains types de documents engendraient des erreurs lors de l'étape de paramétrisation, à cause de caractéristiques inhabituelles.

Parmi les cas les plus parlants, on trouve les poèmes, qui ne comportent aucune ponctuation, et sont considérés par le modèle comme n'ayant qu'une seule phrase. Dès lors, ils obtiennent des valeurs aberrantes pour les variables basées sur la longueur des phrases et, en particulier, **meanNWS**. Ainsi, le texte *a1+/Connexions1/p130.txt* aurait 81,5 mots par phrase. Pire encore, *a2/Tout va bien1/p143.txt* et *a2/Tout va bien1/p152.txt* atteignent respectivement les 257 mots et 282 mots par phrase !

Un autre cas notable est celui des petites annonces, qui biaisent cette fois les variables lexicales. On les rencontre surtout dans les manuels pour débutants. Bien que caractérisé par un lexique simple et concret, ce type d'écrit regorge d'abréviations, comme l'illustre cette

phrase tirée du texte *a1+/Tout va bien1/p81(3).txt* : « Entrée, séj., chamb., coin cuis., s/d'eau, wc, ch. élec., balcon, proximité bus et écoles ». La plupart de ces abréviations n'étant pas très communes dans la langue, elles ne sont pas reconnues par le TreeTagger et se voient considérées comme un mot inconnu, dont la probabilité est dès lors très faible. C'est ce qui explique que ce texte obtienne une valeur de -1007 pour le modèle de langue (**ML1**).

Face à ce genre de problème, il est bien sûr possible de repérer ces données aberrantes une à une et de les éliminer du corpus une fois pour toutes. C'est d'ailleurs l'option que nous avions adoptée précédemment (François, 2009a). Cependant, cette approche souffre de deux défauts. D'une part, elle ne repose pas sur un critère objectif reproductible, puisque la décision de rejeter un texte est prise par un annotateur humain, sur la base de ses observations et de sa connaissance du fonctionnement des prédicteurs. D'autre part, plus le nombre de corpus augmente et plus la taille de ceux-ci croît, moins cette procédure se révèle pratique. C'est pourquoi nous nous sommes résolus à adopter une autre stratégie, basée sur un critère statistique. Celui-ci est simple et consiste à rejeter, au sein de chaque niveau de difficulté, les observations situées au-delà de trois écarts-types de leur moyenne de classe. Ce critère revient à supprimer en moyenne 0.26% des observations pour peu que leur distribution soit normale, ce qui n'est pas toujours le cas. Néanmoins, il est suffisant pour éliminer les textes problématiques tels que les poèmes, les petites annonces ou les textes comprenant trop de noms propres non reconnus par TreeTagger.

Cette technique présente néanmoins aussi un désavantage. Comme elle est adaptée à une distribution univariée, il est nécessaire de l'appliquer plusieurs fois sur un même corpus, afin de traiter chaque variable problématique. Or, à chaque fois qu'elle détecte une observation dont une dimension est considérée comme aberrante, l'observation est supprimée de l'échantillon de test. Dès lors, plus le nombre de variables à traiter augmente, plus la quantité de textes éliminés est susceptible de faire de même.

Afin d'illustrer l'importance de cet effet, un test simple a été réalisé à partir d'un échantillon issu de *Corpus6Cleaned*. La procédure décrite ci-dessus a tout d'abord été appliquée à trois variables différentes : **ML1**, **meanNWS** et **NLML**. Il en ressort que quatre « outliers » seulement ont été détectés. Ensuite, la même procédure a été répliquée pour six variables plus ou moins redondantes (**ML1**, **ML2**, **ML3**, **meanNWS**, **NLML**, et **MeanFL**), avec cette fois comme résultat huit « outliers ». Nous en avons donc conclu qu'il n'était pas possible d'appliquer cette technique sur l'ensemble des variables ayant une distribution normale.

Dès lors, la solution la plus réaliste a consisté à sélectionner un nombre réduit de variables « typiques », représentatives des différents problèmes rencontrés, afin de réaliser la détection des « outliers ». Comme la plupart des prédicteurs issus de la même famille sont sujets aux mêmes particularités textuelles, telles que l'absence de ponctuation pour les mesures

syntaxiques, cette démarche est apparue comme un bon compromis pour opérer une détection optimale des données aberrantes. Les variables sélectionnées sont : **MeanFL**, **PAGough_1063** et **ML3** pour le niveau lexical ; **meanNWS** pour les aspects syntaxiques et **ConcDens** pour la sémantique.

Pour donner un exemple de l'ampleur – limitée – du phénomène, nous rapportons la constitution de huit jeux de données obtenus à partir de nos huit corpus à l'aide de cette procédure automatisée :

– **Test6** : 338 textes, après que 8 « outliers » ont été ôtés ;
– **Test6C** : 343 textes, après le retrait de 8 « outliers » ;
– **Test6E** : 630 textes, après le retrait de 16 « outliers » ;
– **Test6CE** : 625 observations, après l'élimination de 17 « outliers » ;
– **Test9** : 430 textes, après le retrait de 14 « outliers » ;
– **Test9C** : 432 textes, après le retrait de 14 « outliers » ;
– **Test9E** : 940 observations, après que 33 « outliers » ont été ôtés ;
– **Test9CE** : 931 observations, après l'élimination de 37 « outliers ».

Comparaison des corpus

À l'origine, c'est sur ces huit échantillons de tests que nous avions fait porter notre analyse de la capacité de prédiction de 406 variables. Toutefois, cette approche a assez vite fait preuve de deux limitations majeures : l'explosion du nombre de résultats et le manque de fiabilité de ceux-ci.

L'existence de huit jeux de données différents augmente effectivement de manière considérable la quantité de mesures à rapporter, rendant les tableaux de résultats touffus, sans pour autant augmenter considérablement l'intérêt de l'étude. En effet, Guilford (1965, 341) a mis en garde contre la relativité des valeurs absolues des mesures de corrélations, en particulier lorsqu'on compare des corpus dont les propriétés diffèrent : « The size of *r* is very much dependant upon the variability of measured values in the correlated sample. The greater the variability, the higher will be the correlation » [1]. Dès lors, ce ne sont pas tant les valeurs absolues des corrélations intercorpus qui importent que les valeurs relatives, calculées entre les différentes variables sur un même jeu de tests.

Par conséquent, nous avons pris le parti de ne pas présenter les tableaux complets dans le texte, mais de limiter la discussion au meilleur corpus pour chacune des deux échelles de difficulté (6 et 9 niveaux). S'est alors posée la question de savoir comment repérer ces

1. Un constat similaire avait déjà été fait au Chapitre 2.2.5 (Elley, 1969; Klare, 1984) concernant les différences de taille d'effet observées entre différentes études de lisibilité.

deux meilleurs corpus. La procédure que nous avons d'abord imaginée pour atteindre cet objectif compare les différents corpus au moyen de prédicteurs « typiques » des principales dimensions textuelles, à savoir les modèles unigrammes, les variables de la longueur des phrases, la densité du contenu et le type de texte. Autrement dit, elle vise simplement à détecter les corpus produisant les meilleurs résultats pour ces variables.

Cependant, appliquer cette technique revient à prêter le flanc à la critique de Guilford (1965, 341) en comparant les valeurs absolues issues de corpus différents pour les sélectionner. Toutefois, dans notre cas, comme la grande majorité des textes sont communs aux huit corpus, cette remarque apparaît moins cruciale, d'autant qu'elle nous semble pouvoir être ramenée à un problème assez similaire : l'erreur d'échantillonnage.

En effet, comme chacun des échantillons de tests est obtenu à partir d'un corpus plus large, lui-même plus ou moins représentatif de la population à modéliser, il est clair que les valeurs de corrélation obtenues à partir de ces jeux de tests ne correspondent pas exactement à la valeur dans la population. La différence entre le paramètre θ et la statistique T est connue sous le terme d'erreur d'échantillonnage (ε), formellement $\theta = T \pm \varepsilon$. Or, comme la constitution des huit corpus de tests s'effectue de manière aléatoire, ε varie à chaque tirage. Dès lors, un corpus pourra paraître supérieur simplement parce qu'il a profité d'un tirage plus favorable (qui produit une meilleure variabilité des données, par exemple) et non en raison d'une supériorité intrinsèque.

Si on se limite, par exemple, à évaluer les quatre échantillons à 6 classes à l'aide d'une simple procédure reposant sur la comparaison des mesures pour les modèles unigrammes, les variables de la longueur des phrases, la densité du contenu et le type de texte, on obtient les tendances suivantes :

– **modèles unigrammes** : *Test6* < *Test6E* = *Test6CE* < *Test6C* ;
– **variables de longueur des phrases** : *Test6E* < *Test6* < *Test6CE* < *Test6C* ;
– **densité des idées** : *Test6* < *Test6E* < *Test6CE* = *Test6C* ;
– **variables du type de texte** : *Test6* < *Test6E* < *Test6CE* < *Test6C*.

Le corpus d'origine obtient, dans trois cas sur quatre, les moins bonnes performances, ce qui semble entériner les transformations qu'il a subies. Ces dernières ne sont toutefois pas aussi utiles l'une que l'autre. Augmenter la quantité de textes ne semble guère augmenter les corrélations, contrairement à rejeter les textes issus des manuels les plus hétérogènes. D'ailleurs, *Test6CE* obtient même des performances inférieures à *Test6C* à trois reprises, alors qu'il s'agit en théorie du corpus optimal.

Si ce premier tableau apparaît relativement cohérent, on ne peut cependant pas conclure à la supériorité du corpus corrigé (*Corpus6Cleaned*) à partir de ces résultats, car ceux-ci pour-

raient bien n'être dus qu'à un tirage favorable. Dès lors, nous avons implémenté une seconde technique de comparaison, plus fiable. Il s'agit, pour chaque corpus de départ, de générer aléatoirement dix échantillons différents, une approche qui s'apparente au *bootstrapping* (Tufféry, 2007, 396-371). Ensuite, les coefficients de corrélations de chaque variable sont calculés à partir des valeurs obtenues pour chacun de ses 10 échantillons, ce qui réduit $\sqrt{10}$ fois la variabilité des estimateurs obtenus ainsi par rapport à ceux calculés sur un seul jeu de données. Guilford (1965, 348-349) envisage pour ce faire plusieurs techniques de combinaison, mais considère que la moyenne arithmétique est adéquate lorsque les corrélations à synthétiser ne sont pas trop différentes, ce qui est notre cas.

Dès lors, la comparaison des corpus à l'aide des variables « typiques » donne cette fois le motif suivant pour les corpus à six classes [2] :

– **modèles unigrammes** : *Test6* < *Test6E* = *Test6C* < *Test6CE* ;
– **variables de longeur des phrases** : *Test6E* < *Test6* < *Test6CE* < *Test6C* ;
– **densité des idées** : *Test6* < *Test6E* < *Test6CE* < *Test6C* ;
– **variables du type de texte** : *Test6* = *Test6E* < *Test6CE* < *Test6C*.

Ce qui frappe à premier abord est la cohérence entre les résultats obtenus à partir de 10 échantillons et d'un seul. Il semblerait que la quantité d'observations dans les corpus de tests suffise à produire des coefficients de corrélation relativement stables. En effet, on retrouve exactement les mêmes tendances que précédemment : l'infériorité du corpus de base ; l'inefficacité relative des corpus étendus et l'effet nettement plus notable de la tentative de correction de l'hétérogénéité. Ce dernier résultat s'explique en fonction de remarques théoriques de Guilford (1965, 345-346), qui présente l'hétérogénéité de l'échantillon comme une cause de réduction des coefficients de corrélation. Inversement, augmenter l'homogénéité doit donc logiquement conduire à de meilleures corrélations.

Au vu de cette double comparaison de nos huit corpus, il s'avère que les meilleurs sont ceux qui ont été corrigés, c'est-à-dire *Corpus6C* et *Corpus6CE* pour le modèle à 6 classes. Reste à choisir entre ces deux possibilités, sachant que la première obtient de meilleures valeurs pour les variables non lexicales, tandis que la seconde ressort pour les modèles de langue. Étant donné que le choix semble peu porteur de conséquences et qu'il est préférable de calculer les corrélations sur le même corpus que celui qui servira à entraîner les modèles statistiques de la seconde partie, c'est-à-dire *Corpus6CE* et *Corpus9CE*, ce sont ces deux derniers qui ont été retenus pour la suite de nos analyses.

Ce choix opéré, il est désormais possible d'entrer dans le vif du sujet : l'analyse de la capacité prédictive des différentes variables envisagées seules. Il va sans dire que les mesures

2. Les résultats pour les corpus à 9 niveaux sont quasiment similaires et n'ont pas été reproduits ici.

présentées à la suite de cette partie ont été calculées à chaque fois à partir de dix échantillons *bootstrap* et non d'un seul, puisque cette façon de faire produit des estimations plus robustes.

5.2.4 Analyse des variables lexicales

L'analyse des prédicteurs débute avec les variables lexicales. L'ordre de présentation respecte en effet la nomenclature établie au Chapitre 3, cela afin de faciliter la tâche du lecteur désireux de mettre en correspondance considérations théoriques et résultats empiriques, bien que de nombreux liens soient déjà présents dans le texte. Par conséquent, nous passerons successivement en revue : les statistiques de la fréquence lexicale, les absents d'une liste de référence, les modèles unigrammes et n-grammes, les mesures de la diversité lexicale, celles qui évaluent la charge du voisinage orthographique et, finalement, celles qui reposent sur la longueur des mots. La discussion cherche à déterminer le ou les meilleurs prédicteurs au sein de chaque famille et compare également les familles entre elles.

Cependant, non content de rechercher des indices de la lisibilité susceptibles de se révéler intéressants en vue d'une formule de lisibilité pour le FLE, les résultats présentés servent aussi à prendre position vis-à-vis de plusieurs questions théoriques soulevées au Chapitre 3. Celles-ci concernent généralement aussi les variantes au sein d'une même variable, mais la discussion dépasse la simple comparaison des mesures sur base des corrélations pour se faire plus linguistique.

Statistiques de la fréquence lexicale

La première famille de prédicteurs traite des statistiques de la fréquence lexicale. Les tableaux 5.1, 5.2 et 5.3 présentent les résultats des 78 variables de fréquences implémentées, évaluées respectivement au niveau des lemmes, des formes fléchies et des formes fléchies désambiguïsées en fonction de leur catégorie grammaticale. Ces trois tableaux rassemblent un ensemble de résultats riches et touffus, dont nous avons cherché à faire ressortir les grandes tendances, en relation avec les trois interrogations posées dans le Chapitre 3. Celles-ci étaient : quelle est la meilleure statistique ? Quel est le type d'unité linguistique optimal et quelles sont les catégories de mots fournissant les informations les plus précieuses ?

Pour répondre à la première de ces questions, intéressons-nous au premier tableau (5.1), qui traite des mesures basées sur la fréquence des lemmes. On y voit que la moyenne, pourtant privilégiée dans la littérature, se situe, en termes de performances, loin derrière la médiane dans la quasi-totalité des cas. Plus efficaces encore sont les statistiques de dispersion : les 75^{e} et 90^{e} quantiles. En effet, si l'on se base sur les valeurs de r, c'est **X75FL** – le 75^{e} quantile

de la distribution de fréquence des mots du texte – qui se révèle la meilleure variable du lot. Si, par contre, on se repose sur ρ, c'est **X75FLC** – le 75e quantile de la distribution de fréquence des mots de contenus – qui ressort.

	Test6CE				Test9CE			
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
MeanFL	-0.299^{3}	-0.282^{3}	$< 0,001$	0.037	-0.31^{3}	-0.304^{3}	$< 0,001$	0.076
MedianFL	$\mathbf{-0.423}^{3}$	$\mathbf{-0.389}^{3}$	$< 0,001$	0.014	$\mathbf{-0.426}^{3}$	$\mathbf{-0.401}^{3}$	$< 0,001$	0.299
IQFL	-0.006	0.02	$< 0,001$	0.583	-0.005	0.028	$< 0,001$	0.673
X75FL	$\mathbf{-0.45}^{3}$	$\mathbf{-0.579}^{3}$	$< 0,001$	$< 0,001$	$\mathbf{-0.463}^{3}$	$\mathbf{-0.572}^{3}$	$< 0,001$	$< 0,001$
X90FL	-0.445^{3}	-0.562^{3}	$< 0,001$	$< 0,001$	-0.468^{3}	-0.543^{3}	$< 0,001$	$< 0,001$
MeanFLC	-0.327^{3}	-0.297^{3}	$< 0,001$	0.128	-0.321^{3}	-0.297^{3}	$< 0,001$	0.41
MedianFLC	-0.259^{3}	-0.544^{3}	$< 0,001$	0.082	-0.255^{3}	-0.534^{3}	$< 0,001$	0.292
IQFLC	0.352^{3}	0.388^{3}	$< 0,001$	0.087	0.348^{3}	0.382^{3}	$< 0,001$	0.263
X75FLC	-0.32^{3}	$\mathbf{-0.598}^{3}$	$< 0,001$	0.05	-0.399^{3}	$\mathbf{-0.587}^{3}$	$< 0,001$	0.055
X90FLC	-0.343^{3}	-0.531^{3}	$< 0,001$	$< 0,001$	-0.402^{3}	-0.519^{3}	$< 0,001$	$< 0,001$
MeanFLG	-0.078	-0.039	$< 0,001$	0.32	-0.091^{1}	-0.052	$< 0,001$	0.397
MedianFLG	0.053	-0.06	$< 0,001$	0.502	0.032	-0.09^{1}	$< 0,001$	0.522
IQFLG	-0.013	-0.052	$< 0,001$	0.592	-0.058	-0.109^{2}	$< 0,001$	0.618
X75FLG	-0.234^{3}	-0.148^{2}	$< 0,001$	0.007	-0.256^{3}	-0.191^{3}	$< 0,001$	0.003
X90FLG	-0.244^{3}	-0.181^{2}	$< 0,001$	$< 0,001$	-0.299^{3}	-0.243^{3}	$< 0,001$	$< 0,001$
MeanFLN	-0.337^{3}	-0.35^{3}	$< 0,001$	0.093	-0.329^{3}	-0.33^{3}	$< 0,001$	0.413
MedianFLN	-0.326^{3}	-0.461^{3}	$< 0,001$	$< 0,001$	-0.337^{3}	-0.446^{3}	$< 0,001$	0.003
IQFLN	0.286^{3}	0.333^{3}	$< 0,001$	0.165	0.27^{3}	0.299^{3}	$< 0,001$	0.308
X75FLN	-0.318^{3}	-0.504^{3}	$< 0,001$	$< 0,001$	-0.312^{3}	-0.492^{3}	$< 0,001$	$< 0,001$
X90FLN	-0.255^{3}	-0.452^{3}	$< 0,001$	$< 0,001$	-0.248^{3}	-0.454^{3}	$< 0,001$	$< 0,001$
MeanFLV	-0.215^{3}	-0.171^{2}	$< 0,001$	0.03	-0.232^{3}	-0.189^{3}	$< 0,001$	0.198
MedianFLV	-0.362^{3}	$\mathbf{-0.446}^{3}$	$< 0,001$	$< 0,001$	-0.373^{3}	$\mathbf{-0.452}^{3}$	$< 0,001$	$< 0,001$
MeanFLA	-0.201^{3}	-0.133^{2}	$< 0,001$	0.453	-0.191^{3}	-0.119^{2}	$< 0,001$	0.644
MedianFLA	-0.258^{3}	-0.367^{3}	$< 0,001$	0.24	-0.24^{3}	-0.336^{3}	$< 0,001$	0.543
MeanFLC.N	-0.278^{3}	-0.256^{3}	$< 0,001$	0.1	-0.279^{3}	-0.257^{3}	$< 0,001$	0.175
MedianFLC.N	-0.338^{3}	$\mathbf{-0.489}^{3}$	$< 0,001$	0.002	-0.337^{3}	$\mathbf{-0.473}^{3}$	$< 0,001$	0.001

Tableau 5.1 – Variables de fréquences lexicales (lemmes) et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : ^{1} : < 0.05 ; ^{2} : < 0.01 ; ^{3} : < 0.001.

Dès lors, qu'en est-il ? Faut-il se baser sur r ou sur ρ pour la suite de l'analyse, sachant que l'on observe des différences importantes entre les deux types de corrélations pour une partie des variables basées sur les quantiles (médiane comprises) ? Un argument en faveur de ρ apparaît lorsqu'on consulte les p-valeurs du test de linéarité de Guilford, lesquelles sont toujours significatives dans ce cas. Ces valeurs seraient la preuve d'une relation monotonique croissante, mais non linéaire, entre ces prédicteurs et la difficulté des textes. Pour vérifier cette hypothèse, nous avons sélectionné un des 10 échantillons *Test6CE* et y avons étudié plus en détail la distribution bivariée de **X75FLC**, présentée à la Figure 5.3. Celle-ci montre bien que les premiers niveaux comportent beaucoup de textes dont la probabilité du 75e est plus élevée que la moyenne. Ce phénomène influence sévèrement la corrélation de

Pearson, dont la droite d'ajustement est représentée en rouge. D'ailleurs, si l'on compare les moyennes par niveaux, cette relation linéaire monotone apparaît plus clairement encore : A1 : $9,34e-05$; A2 : $5,02e-05$; B1 : $2,52e-05$; B2 : $2,30e-05$; C1 $1,68e-05$; C2 : $1.07e-05$. Nous avons donc tout lieu de préférer Spearman dans ce cas.

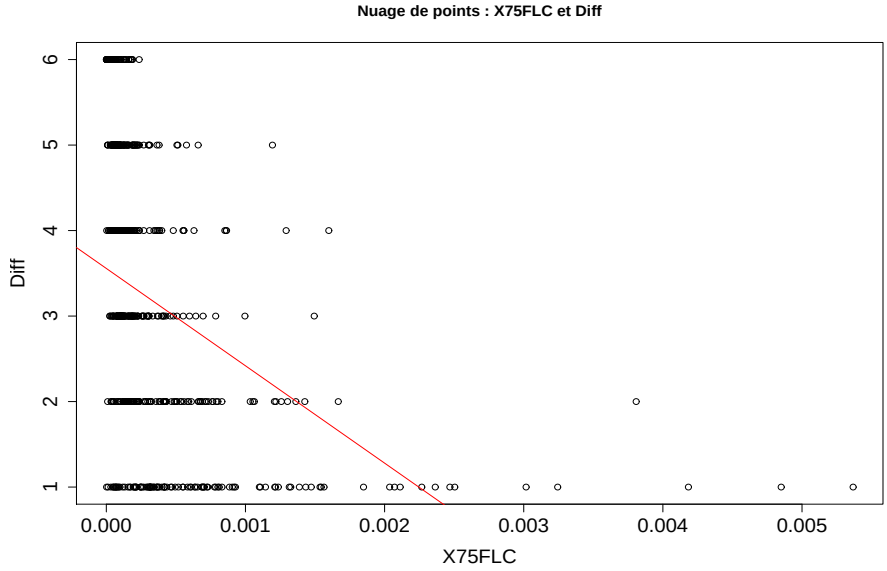


Figure 5.3 – Nuage de points entre X75FLC et la difficulté des textes

Cette question résolue, il est possible d'investiguer le cas des formes fléchies (*cf.* Tableau 5.2) et des formes fléchies désambiguïsées (*cf.* Tableau 5.3). Le constat concernant l'efficacité des différentes statistiques y est globalement le même que pour les lemmes : la médiane est souvent meilleure que la moyenne, en particulier lorsqu'on prend en compte ρ, tandis que les meilleurs prédicteurs sont les quantiles de fin de distribution. C'est tantôt le 90^{e} quantile qui obtient les coefficients les plus élevés, tantôt le 75^{e}. On peut en conclure que ce sont les statistiques de dispersion qui utilisent le mieux l'information de fréquence lexicale pour prédire la difficulté de textes. Or, baser une variable sur un quantile particulier revient à fixer un seuil de fréquence au-delà duquel un mot est considéré comme difficile.

On peut ensuite aller plus loin dans l'analyse et se demander quelle est la meilleure variable parmi les trois tableaux, ce qui revient à se demander quelle est l'unité linguistique optimale. Malheureusement, la réponse est loin d'être simple, car certains indices se comportent mieux lorsqu'ils sont calculés à partir des lemmes ou des formes fléchies. Cependant, les corrélations de Spearman révèlent, qu'en tout cas, le lemme est la moins bonne des trois approches. En ce qui concerne les formes fléchies, il apparaît que la version désambiguïsée favorise les mesures reposant sur le 75^{e} percentile, tandis que la version non désambiguïsée renforce l'effet du 90^{e} percentile. Les deux meilleures sont dès lors **X75FFFDC** et **X90FFFC**, c'est-

	Test6CE				Test9CE			
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
MeanFFF	-0.266^{3}	-0.238^{3}	$< 0,001$	0.013	-0.292^{3}	-0.277^{3}	$< 0,001$	$< 0,001$
MedianFFF	-0.363^{3}	-0.307^{3}	$< 0,001$	0.026	-0.377^{3}	-0.336^{3}	$< 0,001$	0.056
IQFFF	0.137^{1}	0.1	$< 0,001$	0.149	0.169^{3}	0.15^{3}	$< 0,001$	0.08
X75FFF	$\mathbf{-0.372}^{3}$	$\mathbf{-0.564}^{3}$	$< 0,001$	$< 0,001$	$\mathbf{-0.411}^{3}$	$\mathbf{-0.547}^{3}$	$< 0,001$	$< 0,001$
X90FFF	-0.363^{3}	$\mathbf{-0.61}^{3}$	$< 0,001$	$< 0,001$	-0.322^{3}	$\mathbf{-0.584}^{3}$	$< 0,001$	$< 0,001$
MeanFFFC	$\mathbf{-0.358}^{3}$	-0.319^{3}	$< 0,001$	$< 0,001$	$\mathbf{-0.372}^{3}$	-0.341^{3}	$< 0,001$	0.004
MedianFFFC	-0.246^{3}	-0.56^{3}	$< 0,001$	0.099	-0.239^{3}	-0.56^{3}	$< 0,001$	0.083
IQFFFC	0.333^{3}	0.405^{3}	$< 0,001$	0.022	0.343^{3}	0.414^{3}	$< 0,001$	0.054
X75FFFC	-0.253^{3}	-0.624^{3}	$< 0,001$	0.047	-0.331^{3}	-0.619^{3}	$< 0,001$	0.077
X90FFFC	-0.319^{3}	$\mathbf{-0.641}^{3}$	$< 0,001$	$< 0,001$	-0.246^{3}	$\mathbf{-0.628}^{3}$	$< 0,001$	$< 0,001$
MeanFFFG	-0.052	-0.022	$< 0,001$	0.272	-0.073	-0.041	$< 0,001$	0.222
MedianFFFG	-0.19^{3}	-0.161^{3}	$< 0,001$	0.128	-0.2^{3}	-0.179^{3}	$< 0,001$	0.174
IQFFFG	-0.074	-0.08	$< 0,001$	0.505	-0.087^{1}	-0.093^{1}	$< 0,001$	0.573
X75FFFG	-0.212^{3}	-0.146^{2}	$< 0,001$	0.032	-0.238^{3}	-0.189^{3}	$< 0,001$	0.022
X90FFFG	-0.254^{3}	-0.191^{3}	$< 0,001$	0.03	-0.281^{3}	-0.226^{3}	$< 0,001$	0.079
MeanFFFN	-0.254^{3}	-0.257^{3}	$< 0,001$	0.245	-0.22^{3}	-0.234^{3}	$< 0,001$	0.708
MedianFFFN	$\mathbf{-0.359}^{3}$	-0.492^{3}	$< 0,001$	$< 0,001$	$\mathbf{-0.361}^{3}$	-0.469^{3}	$< 0,001$	0.002
IQFFFN	0.321^{3}	0.342^{3}	$< 0,001$	0.193	0.284^{3}	0.307^{3}	$< 0,001$	0.413
X75FFFN	-0.282^{3}	-0.518^{3}	$< 0,001$	$< 0,001$	-0.286^{3}	-0.495^{3}	$< 0,001$	$< 0,001$
X90FFFN	-0.225^{3}	$\mathbf{-0.528}^{3}$	$< 0,001$	0.004	-0.217^{3}	$\mathbf{-0.518}^{3}$	$< 0,001$	$< 0,001$
MeanFFFV	-0.262^{3}	-0.179^{3}	$< 0,001$	$< 0,001$	-0.294^{3}	-0.217^{3}	$< 0,001$	$< 0,001$
MedianFFFV	-0.262^{3}	-0.179^{3}	$< 0,001$	$< 0,001$	-0.294^{3}	-0.217^{3}	$< 0,001$	$< 0,001$
MeanFFFA	-0.186^{3}	-0.088	$< 0,001$	0.303	-0.179^{3}	-0.086^{1}	$< 0,001$	0.594
MedianFFFA	-0.242^{3}	-0.442^{3}	$< 0,001$	0.065	-0.225^{3}	-0.425^{3}	$< 0,001$	0.559
MeanFFFC.N	-0.32^{3}	-0.278^{3}	$< 0,001$	$< 0,001$	-0.338^{3}	-0.304^{3}	$< 0,001$	$< 0,001$
MedianFFFC.N	-0.343^{3}	$\mathbf{-0.509}^{3}$	$< 0,001$	0.005	-0.353^{3}	$\mathbf{-0.514}^{3}$	$< 0,001$	$< 0,001$

Tableau 5.2 – Variables de fréquences lexicales (formes fléchies) et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : ^{1} : < 0.05 ; ^{2} : < 0.01 ; ^{3} : < 0.001.

à-dire deux mesures de la probabilité des mots de contenus.

Cette constatation nous amène à notre dernière interrogation : le type de mots qui pose le plus de problèmes à la lecture, en termes de catégories grammaticales. Selzer (1981, 29) a défendu qu'un indice basé sur les seuls mots lexicaux est préférable, parce que les mots grammaticaux ont tendance à occulter la complexité des termes lexicaux – lesquels sont généralement porteurs de la difficulté – au sein d'une mesure de fréquence. Nos résultats semblent en accord avec cette hypothèse. On observe des ρ supérieurs pour les mesures reposant sur les mots lexicaux que pour celles qui prennent en compte tous les mots, à l'exception de **X90FFFD**. Cet effet est particulièrement sensible pour la médiane, puisque l'on note une différence de $0,155$ entre **MedianFL** et **MedianFLC**, ce qui renforce encore notre conviction en la supériorité des mots lexicaux. En effet, **MedianFL** est probablement influencée par les mots grammaticaux, perdant ainsi de sa capacité de discrimination, tandis que **MedianFLC** capture plus finement les termes complexes, se rapprochant des performances de **X75FLC**.

	Test6CE				Test9CE			
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
MeanFFFD	-0.262^{3}	-0.235^{3}	$< 0,001$	0.006	-0.294^{3}	-0.275^{3}	$< 0,001$	$< 0,001$
MedianFFFD	-0.413^{3}	-0.375^{3}	$< 0,001$	$< 0,001$	-0.434^{3}	-0.4^{3}	$< 0,001$	$< 0,001$
IQFFFD	0.211^{3}	0.157^{2}	$< 0,001$	0.006	0.245^{3}	0.204^{3}	$< 0,001$	$< 0,001$
X75FFFD	-0.388^{3}	-0.596^{3}	$< 0,001$	$< 0,001$	-0.419^{3}	-0.593^{3}	$< 0,001$	$< 0,001$
X90FFFD	$\mathbf{-0.422}^{3}$	$\mathbf{-0.608}^{3}$	$< 0,001$	$< 0,001$	$\mathbf{-0.439}^{3}$	$\mathbf{-0.604}^{3}$	$< 0,001$	$< 0,001$
MeanFFFDC	-0.322^{3}	-0.281^{3}	$< 0,001$	$< 0,001$	-0.341^{3}	-0.307^{3}	$< 0,001$	0.003
MedianFFFDC	-0.219^{3}	-0.564^{3}	$< 0,001$	0.175	-0.207^{3}	-0.569^{3}	$< 0,001$	0.233
IQFFFDC	0.292^{3}	0.417^{3}	$< 0,001$	0.071	0.304^{3}	0.421^{3}	$< 0,001$	0.144
X75FFFDC	-0.296^{2}	$\mathbf{-0.627}^{3}$	$< 0,001$	0.089	-0.367^{3}	$\mathbf{-0.623}^{3}$	$< 0,001$	0.092
X90FFFDC	-0.404^{3}	-0.59^{3}	$< 0,001$	$< 0,001$	-0.381^{3}	-0.589^{3}	$< 0,001$	$< 0,001$
MeanFFFDG	-0.079	-0.05	$< 0,001$	0.178	-0.106^{1}	-0.071	$< 0,001$	0.081
MedianFFFDG	-0.227^{3}	-0.129^{2}	$< 0,001$	0.015	-0.236^{3}	-0.136^{3}	$< 0,001$	0.037
IQFFFDG	-0.098	-0.136^{2}	$< 0,001$	0.581	-0.115^{2}	-0.154^{3}	$< 0,001$	0.486
X75FFFDG	-0.256^{3}	-0.237^{3}	$< 0,001$	$< 0,001$	-0.297^{3}	-0.294^{3}	$< 0,001$	$< 0,001$
X90FFFDG	-0.316^{3}	-0.239^{3}	$< 0,001$	$< 0,001$	-0.353^{3}	-0.291^{3}	$< 0,001$	$< 0,001$
MeanFFFDN	-0.359^{3}	-0.371^{3}	$< 0,001$	0.133	-0.338^{3}	-0.338^{3}	$< 0,001$	0.487
MedianFFFDN	-0.354^{3}	-0.501^{3}	$< 0,001$	$< 0,001$	-0.363^{3}	-0.483^{3}	$< 0,001$	$< 0,001$
IQFFFDN	0.316^{3}	0.342^{3}	$< 0,001$	0.188	0.29^{3}	0.312^{3}	$< 0,001$	0.441
X75FFFDN	-0.288^{3}	$\mathbf{-0.503}^{3}$	$< 0,001$	$< 0,001$	-0.283^{3}	-0.489^{3}	$< 0,001$	$< 0,001$
X90FFFDN	-0.228^{3}	-0.5^{3}	$< 0,001$	0.003	-0.221^{3}	$\mathbf{-0.498}^{3}$	$< 0,001$	$< 0,001$
MeanFFFDV	-0.231^{3}	-0.128^{1}	$< 0,001$	$< 0,001$	-0.267^{3}	-0.168^{3}	$< 0,001$	$< 0,001$
MedianFFFDV	-0.251^{3}	$\mathbf{-0.485}^{3}$	$< 0,001$	$< 0,001$	-0.268^{3}	$\mathbf{-0.51}^{3}$	$< 0,001$	$< 0,001$
MeanFFFDA	-0.18^{3}	-0.083	$< 0,001$	0.308	-0.173^{3}	-0.081^{1}	$< 0,001$	0.563
MedianFFFDA	-0.234^{3}	-0.432^{3}	$< 0,001$	0.073	-0.219^{3}	-0.415^{3}	$< 0,001$	0.553
MeanFFFDC.N	-0.286^{3}	-0.238^{3}	$< 0,001$	$< 0,001$	-0.309^{3}	-0.268^{3}	$< 0,001$	$< 0,001$
MedianFFFDC.N	-0.315^{3}	$\mathbf{-0.51}^{3}$	$< 0,001$	0.016	-0.321^{3}	$\mathbf{-0.519}^{3}$	$< 0,001$	0.003

Tableau 5.3 – Variables de fréquences lexicales (formes fléchies désambiguïsées) et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : ^{1} : < 0.05 ; ^{2} : < 0.01 ; ^{3} : < 0.001.

De même, on peut postuler que la nette différence de corrélation entre **MedianFL** et **X75FL** s'explique par le fait que la majorité des termes difficiles sont des mots lexicaux et que ce dernier indice capture une information sensiblement similaire à **X75FLC** et plus efficace. Cette hypothèse se voit entérinée par les intercorrélations entre ces mesures : alors que **X75FL** et **X75FLC** sont très fortement corrélées ($\rho = 0,89$) – ce qui atteste qu'elles modélisent une information très similaire –, **MedianFL** et **MedianFLC** le sont nettement moins ($\rho = 0,63$). Enfin, une dernière preuve en faveur des mots de contenus plutôt que l'ensemble des termes est, tout simplement, les faibles corrélations obtenues par les variables utilisant uniquement les mots grammaticaux. La meilleure d'entre elles, **X75FFFDG**, obtient un coefficient qui, bien que nettement significatif par rapport à la quantité de données, ne dépasse pas $-0,273$.

Elley (1969) était cependant allé plus loin : elle considérait que seuls les noms devraient faire l'objet d'une évaluation. Nos résultats ne vont toutefois pas dans ce sens, puisque les mesures basées sur les noms sont systématiquement inférieures à celles qui englobent tous

les types de mots lexicaux. Ainsi, non seulement les verbes, adjectifs et adverbes influent eux aussi sur la difficulté des textes, mais il s'avère même que les médianes des mesures qui ne prennent que l'une de ces catégories en compte (**MedianFLV**, **MedianFLA**) ou leur ensemble (**MedianFLC.N**) surpasse **MedianFLN**. Nous en venons donc à rejeter l'hypothèse d'Elley (1969). D'après nos données, du moins, les meilleures variables de fréquences sont donc **X75FFFDC** et **X90FFFC**. Notons cependant que ces deux indices sont fortement intercorrélés ($\rho = 0,88$) et qu'il est donc probable qu'un seul d'entre eux suffise au sein d'une formule de lisibilité.

Les absents d'une liste de référence

Le pourcentage de mots absents d'une liste de référence a depuis longtemps fait ses preuves dans la littérature (*cf.* Chapitre 3.2.2). On peut donc s'attendre à ce que les prédicteurs de cette famille se révèlent efficaces et c'est effectivement le cas, puisque la corrélation de Spearman de la meilleure variable, **PA_Alterego1a**, atteint $0,672$. Par contre, il est plus étonnant de constater que **PA_Alterego1a** n'est rien moins que la meilleure variable de notre ensemble, toutes dimensions textuelles confondues, surpassant de loin les indices de la fréquence lexicale, et même les modèles unigrammes.

Une telle observation apparaît quelque peu inattendue au regard de la littérature qui a souvent considéré le nombre d'absents comme une estimation des aspects fréquentiels (Lively et Pressey, 1923) et que Elley (1969) n'a pas hésité à qualifier d'« indice grossier ». Néanmoins, certains faits déjà rapportés précédemment vont dans le même sens que nos résultats. Rappelons qu'Henry (1975, 75) avait déjà constaté que, pour le français, cet indice avait un pouvoir discrimant égal à celui des prédicteurs basés sur la fréquence des mots. Par ailleurs, le pourcentage d'absents se retrouve dans la formule de Dale et Chall (1948a), reconnue comme l'une des meilleures. Nos propres analyses confirment donc cette capacité discriminante exceptionnelle, du moins en ce qui concerne le FLE.

Au-delà de la simple évaluation des corrélations, il est intéressant de voir comment les résultats de l'ensemble des prédicteurs de cette famille, présentés dans le Tableau 5.4, éclairent les questions que nous avions soulevées dans le Chapitre 5. La première d'entre-elles concerne l'évolution de la capacité de discrimination en fonction de la taille de la liste, mise en évidence tant par Dale et Chall (1948a) que Harris et Jacobson (1974). À ce niveau, il apparaît clairement qu'une liste courte est préférable pour le FLE. Pour la liste de Gougenheim, les corrélations optimales se situent à 2000 mots lorsqu'on considère l'ensemble des formes fléchies du texte et autour de 3000 mots pour les « types ». Similairement, la liste courte (802 mots) d'*Alter Ego* se comporte mieux que celle qui comprend 1803 mots. Ce résultat n'est guère étonnant, étant donné que les extraits issus de manuels pour débutants sont simples et

	Test6CE				Test9CE			
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
PAGoug_1063	0.578^{3}	0.577^{3}	$< 0,001$	0.008	0.562^{3}	0.574^{3}	$< 0,001$	0.235
PAGoug_2000	$\mathbf{0.593}^{3}$	$\mathbf{0.597}^{3}$	$< 0,001$	0.017	$\mathbf{0.574}^{3}$	$\mathbf{0.588}^{3}$	$< 0,001$	0.313
PAGoug_3000	0.584^{3}	0.591^{3}	0.015	0.018	0.567^{3}	0.586^{3}	$< 0,001$	0.317
PAGoug_4000	0.578^{3}	0.585^{3}	0.036	0.031	0.56^{3}	0.578^{3}	$< 0,001$	0.327
PAGoug_5000	0.567^{3}	0.578^{3}	0.029	0.132	0.543^{3}	0.565^{3}	$< 0,001$	0.203
PAGoug_6000	0.549^{3}	0.563^{3}	0.054	0.247	0.515^{3}	0.54^{3}	$< 0,001$	0.109
PAGoug_7000	0.525^{3}	0.543^{3}	0.024	0.267	0.486^{3}	0.517^{3}	$< 0,001$	0.077
PAGoug_8000	0.514^{3}	0.53^{3}	0.022	0.356	0.471^{3}	0.502^{3}	$< 0,001$	0.078
PAGoug_8775	0.502^{3}	0.519^{3}	0.016	0.374	0.451^{3}	0.486^{3}	$< 0,001$	0.032
PAGougU_1063	0.523^{3}	0.518^{3}	$< 0,001$	$< 0,001$	0.529^{3}	0.541^{3}	$< 0,001$	0.019
PAGougU_2000	0.542^{3}	0.547^{3}	0.002	$< 0,001$	0.546^{3}	0.564^{3}	$< 0,001$	0.035
PAGougU_3000	$\mathbf{0.549}^{3}$	$\mathbf{0.555}^{3}$	0.054	$< 0,001$	$\mathbf{0.549}^{3}$	$\mathbf{0.571}^{3}$	0.002	0.116
PAGougU_4000	0.544^{3}	$\mathbf{0.556}^{3}$	0.089	0.001	0.548^{3}	$\mathbf{0.572}^{3}$	$< 0,001$	0.193
PAGougU_5000	0.538^{3}	$\mathbf{0.557}^{3}$	0.016	0.012	0.539^{3}	0.569^{3}	$< 0,001$	0.339
PAGougU_6000	0.522^{3}	0.544^{3}	0.007	0.112	0.515^{3}	0.548^{3}	$< 0,001$	0.448
PAGougU_7000	0.503^{3}	0.528^{3}	0.003	0.177	0.489^{3}	0.526^{3}	$< 0,001$	0.45
PAGougU_8000	0.491^{3}	0.518^{3}	0.006	0.253	0.475^{3}	0.514^{3}	$< 0,001$	0.494
PAGougU_8775	0.484^{3}	0.507^{3}	0.006	0.319	0.46^{3}	0.497^{3}	$< 0,001$	0.356
PA_Alterego1a	$\mathbf{0.657}^{3}$	$\mathbf{0.652}^{3}$	$< 0,001$	$< 0,001$	$\mathbf{0.668}^{3}$	$\mathbf{0.672}^{3}$	$< 0,001$	0.002
PAU_Alterego1a	$\mathbf{0.093}$	$\mathbf{0.105}$	$< 0,001$	0.002	$\mathbf{0.129}^{2}$	$\mathbf{0.142}^{3}$	$< 0,001$	0.004
PA_Alterego1b	0.649^{3}	0.636^{3}	$< 0,001$	$< 0,001$	0.659^{3}	0.656^{3}	$< 0,001$	0.001
PAU_Alterego1b	0.569^{3}	0.55^{3}	$< 0,001$	$< 0,001$	0.603^{3}	0.601^{3}	$< 0,001$	$< 0,001$
PA_Alterego2a	0.621^{3}	0.623^{3}	$< 0,001$	0.009	0.604^{3}	0.616^{3}	$< 0,001$	0.248
PAU_Alterego2a	0.03	0.057	$< 0,001$	0.028	0.045	0.072	$< 0,001$	0.016
PA_Alterego2b	0.631^{3}	0.627^{3}	$< 0,001$	0.005	0.618^{3}	0.623^{3}	$< 0,001$	0.198
PAU_Alterego2b	0.563^{3}	0.556^{3}	0.009	$< 0,001$	0.573^{3}	0.585^{3}	0.022	0.001

Tableau 5.4 – Pourcentage d'absents et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : ^{1} : < 0.05 ; ^{2} : < 0.01 ; ^{3} : < 0.001.

probablement assimilables à des textes pour des jeunes primaires, à propos desquels Harris et Jacobson (1974) montraient qu'une liste courte est plus adaptée.

Pour élucider la différence de taille optimale entre les listes de Gougenheim et d'*Alter-Ego*, rappelons que cette dernière ne comprend que des mots lexicaux. Il est donc logique qu'elle nécessite moins de termes pour atteindre le palier des mots « difficiles ». Par contre, on notera que, pour la liste courte, le fait de se limiter aux termes lexicaux améliore légèrement les corrélations par rapport à **PA_Alterego1b**, où tous les mots des textes ont été considérés suite à l'ajout d'une mini-liste de 62 termes grammaticaux. Par contre, l'effet s'inverse dans le cas de la liste longue (**PA_Alterego2a** et **PA_Alterego2b**).

Il n'est toutefois pas possible de considérer la variante **b** des listes d'*Alter-Ego* comme réellement équivalente à une approche où tous les types de mots sont pris en compte dans la liste. Dès lors, nos expériences ne permettent pas de déterminer si la différence entre **PA_Alterego1a** et **PAGoug_2000** provient de cette limitation aux termes lexicaux, de la modernité de la liste ou encore d'une plus grande proximité avec le vocabulaire du corpus.

Ce qui est certain, du moins, c'est que le fait d'employer une liste fournie à la fin d'un seul manuel A1 permet de discriminer efficacement des textes issus d'une vingtaine d'ouvrages, répartis sur l'ensemble des 6 ou 9 niveaux du CECR.

Reste toutefois à clarifier un troisième point : est-il plus opérant de considérer les formes fléchies ou les « types » dans le décompte de la proportion d'absents ? Autrement dit, est-ce que l'effet d'un mot absent ne doit être pris en compte qu'une seule fois ou à chacune de ses occurrences ? Sur ce point, nos résultats sont en contradiction avec Harris et Jacobson (1974) et Nyman *et al.* (1961), lesquels observaient la supériorité du comptage des « types » seuls. Or, sur nos propres données, les indices basés sur les formes fléchies surpassent systématiquement ceux reposant sur les « types », ce qui montre que la répétition de mots complexes n'aide pas vraiment à réduire la difficulté d'un texte.

Notons enfin que, comme le signe de la corrélation est positif, la difficulté des textes augmente logiquement en proportion du nombre d'absents dans un texte. Par ailleurs, toutes les variables peuvent être considérées comme non normales et si la majorité de celles basées sur la liste de Gougenheim semble entretenir une relation linéaire avec la difficulté, ce n'est pas le cas de celles dépendant de la liste d'*Alter-Ego*. Il s'agit là d'un résultat que nous ne nous expliquons pas et qui n'a pas fait l'objet d'investigation plus poussées, car il sort du focus de notre étude. Il est peut être simplement dû à des effets aléatoires. Quoi qu'il en soit, il ressort nettement de nos analyses que les deux variantes les plus efficaces dans cette famille sont **PAGoug_2000** et **PA_Alterego1a**.

Les modèles unigrammes

Avec les modèles unigrammes, notre analyse entre dans la problématique de la vraisemblance des textes, dont ce modèle constitue l'approximation la plus simple, sans en être moins efficace pour autant. Le Tableau 5.5, où sont repris les résultats obtenus par les trois variantes (**ML1**, **ML2** et **ML3**.), confirme nos résultats précédents (François, 2009a,b) et atteste de la corrélation élevée qu'entretient cet indice lexical avec la difficulté.

	Test6CE				Test9CE			
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
ML1	-0.521^3	-0.522^3	0.031	0.076	-0.498^3	-0.513^3	0.038	0.538
ML2	-0.53^3	-0.52^3	$< 0,001$	0.013	-0.508^3	-0.507^3	0.022	0.229
ML3	$\mathbf{-0.56}^3$	$\mathbf{-0.546}^3$	$< 0,001$	$< 0,001$	$\mathbf{-0.556}^3$	$\mathbf{-0.552}^3$	$< 0,001$	0.026

Tableau 5.5 – Modèles unigrammes et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : $^1 : < 0.05$; $^2 : < 0.01$; $^3 : < 0.001$.

On observe que les résultats sont très cohérents entre les deux corpus : l'approche par les lemmes (**ML1**) est systématiquement surpassée par celle qui repose sur les formes fléchies (**ML2**), elle-même moins pertinente que **ML3**, basée sur les formes fléchies désambiguïsées d'après leur catégorie grammaticale. Autrement dit, l'efficacité augmente de pair avec une prise en compte accrue des différences entre les mots : d'une approche qui confond les probabilités des formes fléchies d'un même lemme, on va vers une autre qui cherche à modéliser des différences de fréquence au sein d'une même forme graphique.

Bien que ce portrait apparaisse séduisant, il pourrait très bien être biaisé. En effet, rappelons que le retrait des « outliers » a été opéré uniquement sur **ML3**. Or, supprimer des « outliers » améliore généralement les corrélations. C'est pourquoi une expérience supplémentaire a été menée afin d'éviter ce biais. Tout simplement, les mesures présentées du Tableau 5.5 ont été recalculées sur des échantillons où **ML1**, **ML2** et **ML3** – et uniquement ces trois variables – ont été utilisées dans la procédure de suppression des données aberrantes. Les nouvelles mesures sont rapportées dans le Tableau 5.6. Elles révèlent exactement la même configuration que précédemment, ce qui nous permet d'affirmer avec plus de force que, pour le français, il est préférable de modéliser la vraisemblance lexicale par modèle unigramme sur base des formes fléchies désambiguïsées.

	Test6CE				Test9CE			
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
ML1	-0.532^3	-0.537^3	0.009	0.132	-0.505^3	-0.515^3	0.027	0.674
ML2	-0.543^3	-0.536^3	0.002	0.06	-0.518^3	-0.512^3	$< 0,001$	0.273
ML3	$\mathbf{-0.572}^3$	-0.566^3	$< 0,001$	0.006	$\mathbf{-0.559}^3$	-0.555^3	$< 0,001$	0.013

Tableau 5.6 – Modèles unigrammes et difficulté : résultats de l'expérience avec retrait des « outliers » dans les trois variables.

Il s'agit là d'une avancée par rapport à nos expériences précédentes (François, 2009a), où nous n'avions pas observé de différences significatives entre les trois facteurs. Dans ce cas-ci, nous avons appliqué un test T de Student (Howell, 2008, 199) aux différences entre les corrélations moyennes obtenues et il apparaît que (1) pour *Test6CE*, **ML3** est effectivement significativement différente de **ML1** ($T = 4,85; p < 0.0001$) comme de **ML2** ($T = 4,31; p = 0,0002$) ; (2) pour *Test9CE*, la différence est encore plus significative entre **ML3** et **ML1** ($T = 6,55; p < 0.0001$), ainsi qu'entre **ML3** et **ML2** ($T = 6,55; p < 0,0001$).

À coté de cette supériorité de **ML3**, il est également intéressant de remarquer que (1) comme attendu, plus la difficulté des textes augmente et moins le vocabulaire des textes est fréquent (corrélation négative) ; (2) les variables unigrammes montrent une certaine tendance à la linéarité (avec des réserves concernant **ML3**), laquelle est confirmée par le fait que les corrélations de Spearman ne sont pas meilleures que celles de Pearson ; (3) enfin, une fois encore,

le test de Shapiro-Wilk conduit à rejeter l'hypothèse de normalité des trois variables.

Les modèles n-grammes

On peut se demander si cette efficacité des modèles unigrammes se reproduit pour les modèles d'ordre supérieur [3]. Nous avions ainsi prévu d'expérimenter huit variables pour chacun des ordres des n-grammes (2 à 5) référencés dans nos deux références : Google et *Le Soir*. Ce projet ambitieux a toutefois été très vite revu à la baisse, après l'analyse des variables basées sur les bigrammes (*cf.* Tableau 5.7). En effet, il est apparu que la proportion élevée d'événements inconnus apparaissant dans les textes à évaluer constituait déjà une sérieuse entrave à l'efficacité des variables de cet ordre.

Ainsi, la proportion d'événements inconnus a été calculée sur l'un des 10 échantillons tirés de *Corpus6CE* : elle atteint respectivement 61% pour la référence *Le Soir* et 36% pour celle basée sur Google. Cela signifie qu'un pourcentage équivalent de n-grammes des textes évalués se sont vus attribuer la même probabilité : celle réservée aux événements inconnus. Or, pour peu que ce taux dépasse les 50%, il en découle que la quasi-totalité des 75^{e} et 90^{e} percentiles, ainsi que, dans une moindre mesure, un grand nombre de médianes des textes reçoivent aussi cette probabilité. Il est dès lors évident que la capacité de discrimination des variables qui en découlent est nulle, comme le confirment les corrélations de **NGprob90.LS**, **NGprob90.G**, **NGprob75.LS** et même **medianNGProb.LS**. C'est pourquoi il n'a pas été jugé utile de pousser l'analyse au-delà du bigramme.

	Test6CE				Test9CE			
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
normTLProb.LS	-0.072	-0.022	$< 0,001$	0.177	-0.062	-0.011	$< 0,001$	0.566
normTLProb.G	-0.084	-0.011	$< 0,001$	0.31	-0.069	0.02	$< 0,001$	0.499
MPCond.LS	$\mathbf{-0.196}^{3}$	-0.113^{1}	$< 0,001$	$< 0,001$	$\mathbf{-0.222}^{3}$	-0.124^{2}	$< 0,001$	$< 0,001$
MPCond.G	$\mathbf{-0.255}^{3}$	-0.186^{3}	$< 0,001$	$< 0,001$	$\mathbf{-0.277}^{3}$	-0.198^{3}	$< 0,001$	$< 0,001$
MedPCond.LS	-0.08	-0.029	$< 0,001$	0.113	-0.049	0.003	$< 0,001$	0.138
MedPCond.G	-0.151^{3}	-0.056	$< 0,001$	0.386	-0.123^{3}	-0.027	$< 0,001$	0.42
meanNGProb.LS	$\mathbf{0.28}^{3}$	$\mathbf{0.303}^{3}$	$< 0,001$	0.162	$\mathbf{-0.267}^{3}$	-0.13^{2}	$< 0,001$	0.342
meanNGProb.G	$\mathbf{0.382}^{3}$	$\mathbf{0.407}^{3}$	0.011	0.05	$\mathbf{-0.244}^{3}$	-0.104^{1}	$< 0,001$	0.417
gmeanNGProb.LS	0.091	0.128^{2}	$< 0,001$	0.077	0.031	0.087^{1}	$< 0,001$	0.383
gmeanNGProb.G	-0.009	0.118^{1}	$< 0,001$	0.387	-0.014	0.127^{3}	$< 0,001$	0.676
medianNGProb.LS	-0.088	-0.079	$< 0,001$	0.42	-0.081^{1}	-0.065	$< 0,001$	0.676
medianNGProb.G	-0.002	0.114^{1}	$< 0,001$	0.39	0.011	0.135^{3}	$< 0,001$	0.433
NGprob75.LS	0.058	0.058	/	0.64	/	/	/	/
NGprob75.G	-0.096	-0.107^{1}	$< 0,001$	0.522	-0.081^{1}	-0.071	$< 0,001$	0.771
NGprob90.LS	0.058	0.058	/	0.64	/	/	/	/
NGprob90.G	-0.072	-0.065	$< 0,001$	0.523	-0.061	-0.051	$< 0,001$	0.674

Tableau 5.7 – Variables de n-grammes et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : $^{1} : < 0.05$; $^{2} : < 0.01$; $^{3} : < 0.001$.

3. Soulignons que l'analyse détaillée dans cette section a fait l'objet de deux publications distinctes, réalisées en collaboration avec Patrick Watrin (François et Watrin, 2011a,b).

En ce qui concerne les autres variables, une première constatation surprenante est l'inefficacité des prédicteurs basés sur un modèle bigramme classique : le r de **normTLProb** vaut -0.072 pour *Le Soir* et -0.084 pour Google. En effet, ce résultat apparaît contradictoire en regard des travaux sur l'anglais. Schwarm et Ostendorf (2005), par exemple, disent employer des modèles n-grammes avec succès, même s'ils ne rapportent pas de corrélation individuelle pour cette variable et si leurs bonnes performances globales sont obtenues à l'aide de plusieurs prédicteurs. Étant donné que le niveau unigramme se comporte très bien, il nous faut conclure soit à un problème d'estimation des probabilités au niveau des deux références, soit remettre en doute l'hypothèse qui fonde l'emploi d'un modèle n-gramme en lisibilité, à savoir que la probabilité d'un mot étant donné son historique est liée à sa difficulté de reconnaissance et de compréhension. Ce dernier aspect requerrait cependant des expérimentations dans le cadre de modèles explicatifs et non prédictifs.

Ces résultats sont d'autant plus étonnants que la moyenne des probabilités des bigrammes du texte apparaît quant à elle largement significative ($r = 0.28$ pour **meanNGProb.LS** et $r = 0.382$ pour **meanNGProb.G**)[4]. Par contre, la moyenne géométrique de ces mêmes unités (**gmeanNGProb.LS** et **gmeanNGProb.G**) présente à nouveau des corrélations non significatives (respectivement $0,091$ et $-0,009$) pour r et légèrement significatives pour ρ (respectivement $0,128$ et $0,118$).

Confronté à ces résultats, il nous semble pouvoir conclure à un effet négatif de la multiplication par rapport à l'addition des probabilités des n-grammes. D'ailleurs, cette même situation se reproduit dans le cas des probabilités conditionnelles, où **MPCond** se comporte mieux que **normTLProb**. Cet effet, que nous n'avons trouvé nulle part dans la littérature mériterait d'être investigué davantage.

	meanNGProb.LS	meanNGProb.G	MPCond.LS	MPCond.G	ML3	NMP	MeanPColl
meanNGProb.LS	1	$\mathbf{0,12}$	$0,44$	$0,01$	$-0,06$	$0,24$	$0,06$
meanNGProb.G	$0,12$	1	$-0,1$	$0,07$	$-0,2$	$\mathbf{0,35}$	$-0,04$
MPCond.LS	$0,44$	$-0,1$	1	$\mathbf{0,57}$	$0,07$	$-0,1$	$0,11$
MPCond.G	$0,01$	$0,07$	$0,57$	1	$0,13$	$-0,14$	$0,1$

Tableau 5.8 – Corrélations croisées (ρ de Spearman) entre les quatre meilleures variables n-grammes et quelques variables d'intérêt, calculées sur l'un des échantillons issus de Corpus6CE.

Une autre constatation d'intérêt apparaît à la lecture du Tableau 5.8, qui présente quelques corrélations croisées entre facteurs. On y découvre ainsi que les mêmes variables, calculées à partir de références différentes, ne sont que très peu corrélées. Ce résultat surprenant témoigne d'une grande dissimilarité entre ces deux bases, tant au niveau des n-grammes

4. Nous avons d'ailleurs montré par ailleurs (François et Watrin, 2011a) que cette variable est plus informative que **MeanPColl**, une variable pourtant plus motivée linguistiquement, qui calcule la moyenne des probabilités des UPs.

présents que de leurs fréquences. Par ailleurs, l'association de ces quatre facteurs avec les prédicteurs lexicaux est quasi nulle ; il semblerait que l'information apportée par les modèles n-grammes soit davantage de nature syntaxique.

En résumé, les variables basées sur les n-grammes ne semblent guère informatives *a priori*. Certes, la meilleure d'entre-elles (**meanNGProb.G**) atteint tout de même un ρ de $0,407$ sur le *Corpus6CE*, mais ses performances se dégradent sensiblement sur le *Corpus9CE*. De plus, ce résultat reste inférieur à nombre d'autres prédicteurs lexicaux et, au vu des ressources à mettre en place pour l'obtenir, ces efforts computationnels apparaissent démesurés par rapport aux résultats. Notons toutefois que ce type d'indices, situés entre le lexique et la syntaxe, apportent bien une information nouvelle, comme le révèle le Tableau 5.8. Il est donc possible qu'ils se révèlent intéressants lors de l'étape de modélisation de la formule.

Les mesures de la diversité lexicale

En ce qui concerne les mesures de la diversité lexicale, le tableau est là aussi assez sombre. Déjà, lors d'expériences préalables (François, 2009a), nous avions conclu au faible apport informatif du TTR classique et l'avions exclu de notre première formule. Nous avons alors soupçonné un biais relatif à la longueur des textes : les documents les plus longs (et souvent les plus ardus) obtiennent des valeurs plus basses, ce qui contrebalancerait le fait que les textes complexes ont normalement un TTR plus élevé. C'est pourquoi une version normalisée de cet indice a été implémentée dans l'espoir de passer outre cette faiblesse.

	Test6CE				Test9CE			
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
TTR_W	-0.283^3	-0.284^3	0.008	0.004	-0.251^3	-0.253^3	0.005	0.005
TTR_L	-0.269^3	-0.26^3	$< 0,001$	0.009	-0.25^3	-0.241^3	$< 0,001$	0.025
NormTTR_W	0.125^2	0.086	$< 0,001$	0.015	0.171^3	0.13^3	$< 0,001$	$< 0,001$
NormTTR_L	0.092	0.08	$< 0,001$	0.116	0.127^2	0.107^2	$< 0,001$	0.013

Tableau 5.9 – Mesures de la diversité lexicale et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : $^1 : < 0.05$; $^2 : < 0.01$; $^3 : < 0.001$.

Il ressort des résultats présentés dans le Tableau 5.9 que la normalisation n'a pas amélioré l'efficacité de cette variable. Au contraire, pour les 10 échantillons de *Corpus6CE*, on passe de -0.284 pour **TTR_W** (une corrélation qui, bien que peu élevée, reste nettement significative) à 0.086 pour **NormTTR_W**. Rappelons toutefois que la normalisation opérée s'appuie sur un paramètre n, à savoir la longueur, en nombre de mots, des fragments sur lesquels un TTR moyen est calculé (*cf.* Chapitre 3). La valeur employée pour le calcul des variables rapportées dans le Tableau 5.9 est 100. Nous avons également testé l'ensemble des valeurs

de n comprises entre 100 et 1000 (par pas de 50). La Figure 5.4 montre, pour chacune de ces valeurs de n, les corrélations de Spearman moyennes, estimées sur 10 échantillons issus de *Corpus6CE*. On s'aperçoit que si la corrélation augmente conjointement à n, elle tend en réalité vers le score de **TTR_W**. Cela est simplement dû au fait que presque aucun texte de notre corpus ne dépasse les 1000 mots et que considérer des fragments de 1000 mots revient à calculer le TTR classique. Par conséquent, nous avons conservé $n = 100$ pour la suite de notre étude, afin de conserver un sens à l'opération de normalisation.

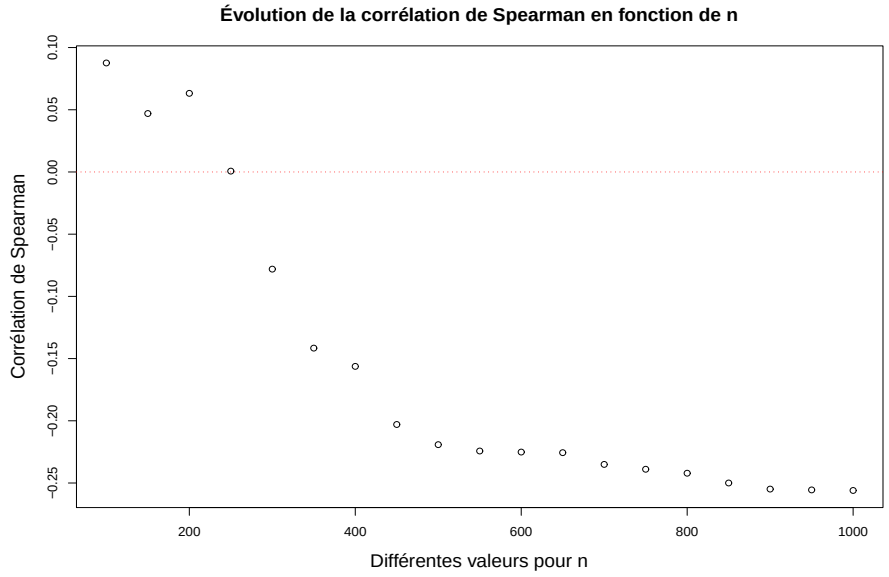


*Figure 5.4 – Variations du coefficient de corrélation de Spearman pour **NormTTR_L** en fonction de n.*

Ce type de variables apparaît donc peu intéressant pour notre formule de lisibilité pour le FLE. Il faut dire qu'il s'agit d'un indice n'ayant jamais fait l'unanimité dans la littérature (*cf.* Chapitre 3), à la différence des mesures basées sur la fréquence. Néanmoins, cela n'explique toujours pas pourquoi, alors qu'il est fréquemment employé en ALS (notamment dans le programme Coh-Metrix), il se révèle un aussi médiocre prédicteur. Il semblerait que, tout simplement, l'hypothèse selon laquelle les textes de FLE plus simples aient un TTR plus bas soit peu réalisée dans notre corpus, comme le révèlent les valeurs moyennes par niveaux (cf. tableau 5.10). On y découvre un profil en dents de scie, où la différence entre A1 et C2 n'est que de $2,8\%$. Notons enfin que, de manière générale, le fait de travailler avec les lemmes ne change rien à ce constat et tend même à l'aggraver. Il faut donc conclure à la faible capacité prédictive des variables basées sur le TTR.

Niveaux	A1	A2	B1	B2	C1	C2
Moyennes	$62,7\%$	$67,2\%$	$64,8\%$	$61,3\%$	$56,7\%$	$59,9\%$

Tableau 5.10 – Moyennes du TTR par niveau de difficulté, calculées sur un échantillon du ***Corpus6CE****.*

Le voisinage orthographique

Comme nous y avons déjà insisté, le voisinage orthographique n'a jamais fait l'objet d'expérimentations en lisibilité, alors que son influence sur la lecture a été confirmée théoriquement. Les travaux cités au Chapitre 3 font état d'un effet d'inhibition pour le français lorsqu'un mot à reconnaître comporte un voisin orthographique plus fréquent. Sans préjuger du bien fondé de cette affirmation, il apparaît que son utilité pratique en lisibilité est faible. En effet, le nombre moyen de termes ayant un voisin plus fréquent (**mean_neighMoreFreq**) obtient une corrélation non significative ($\rho = -0,037$) pour le *Corpus6CE*. Ce n'est guère étonnant, puisque lorsqu'on observe les données de plus près, on s'aperçoit que ce nombre reste stable entre les niveaux, passant de $40,2\%$ pour A1 à $39,6\%$ pour C2. On peut donc en conclure que si les mots possédant des voisins plus fréquents sont reconnus plus difficilement, leur distribution ne varie pas nécessairement au sein des différents niveaux de difficulté des textes.

	Test6CE				Test9CE			
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
mean_nbNeigh	-0.086	-0.043	$< 0,001$	0.182	-0.096^{1}	-0.05	$< 0,001$	0.194
med_nbNeigh	-0.075	-0.07	$< 0,001$	0.358	-0.084^{1}	-0.072	$< 0,001$	0.353
nbNeigh_75	0.089	0.126^{1}	$< 0,001$	0.465	0.078	0.124^{2}	$< 0,001$	0.67
nbNeigh_90	-0.194^{3}	-0.099^{1}	$< 0,001$	0.099	-0.184^{3}	-0.112^{1}	$< 0,001$	0.271
mean_neighMoreFreq	-0.031	$\mathbf{-0.037}$	$< 0,001$	0.429	-0.019	$\mathbf{-0.031}$	$< 0,001$	0.178
mean_nbNeighMoreFreq	-0.165^{3}	-0.068	$< 0,001$	0.367	-0.143^{3}	-0.041	$< 0,001$	0.555
med_nbNeighMoreFreq	-0.219^{3}	$\mathbf{-0.229}^{3}$	$< 0,001$	0.359	-0.175^{3}	-0.186^{3}	$< 0,001$	0.386
nbNeighMoreFreq_75	-0.104^{1}	-0.124^{2}	$< 0,001$	0.422	-0.075^{1}	-0.093^{1}	$< 0,001$	0.601
nbNeighMoreFreq_90	-0.13^{2}	-0.192^{3}	$< 0,001$	0.44	-0.108^{2}	-0.144^{3}	$< 0,001$	0.59
mean_freqCumNeigh	0.173^{3}	0.19^{3}	$< 0,001$	0.571	0.172^{3}	0.185^{3}	$< 0,001$	0.27
med_freqCumNeigh	-0.083	0.022	$< 0,001$	0.317	-0.105^{2}	-0.003	$< 0,001$	0.111
freqCumNeigh_75	-0.151^{1}	$\mathbf{-0.306}^{3}$	$< 0,001$	0.21	-0.154^{3}	$\mathbf{-0.292}^{3}$	$< 0,001$	0.164
freqCumNeigh_90	-0.145^{2}	-0.274^{3}	$< 0,001$	0.156	-0.158^{3}	-0.245^{3}	$< 0,001$	0.061

Tableau 5.11 – Mesures du voisinage orthographique et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : ^{1} : < 0.05 ; ^{2} : < 0.01 ; ^{3} : < 0.001.

Confronté à ce constat, on pourrait penser, qu'une fois de plus, c'est la moyenne qui se révèle inadéquate. Il est vrai que les autres statistiques, à savoir la médiane et les 75e et 90e percentiles obtiennent des résultats supérieurs. Néanmoins, le coefficient de la meilleure d'entre elles ne dépasse pas $-0,306$, ce qui est loin de faire de **freqCumNeigh_75** l'une de nos va-

riables les plus informatives. On pourrait également soutenir que ces résultats s'expliquent par un problème de couverture de la liste distribuée dans le cadre du projet Lexique3. Ce n'est toutefois pas le cas, puisque la couverture lexicale oscille entre $81,6\%$ pour A1 et $77,4\%$ pour C2, ce qui représente déjà un score très correct.

Ainsi, en définitive, il est fort peu probable que cette famille de variables, dont les plus associées à la difficulté sont celles basées sur la fréquence cumulée de tous les voisins, se révèle utile pour la seconde partie de notre étude. À moins que l'information qu'elle apporte ne soit de nature différente de celle des autres prédicteurs lexicaux, ce qui semble être en partie le cas. Nous avons en effet mesuré la corrélation croisée entre **freqCumNeigh_75** et deux des meilleures variables lexicales découvertes jusqu'à présent : **ML3** et **PA_Alterego1a** et avons obtenu une association de taille moyenne dans les deux cas (respectivement $\rho = 0,47$ et $\rho = -0,4$). En comparaison, **ML3** et **PA_Alterego1a** sont nettement plus redondants, puisque leur ρ atteint $-0,66$.

La longueur des mots

Faisant partie des variables lexicales les plus populaires en lisibilité depuis les travaux de Smith (1961), la longueur moyenne des mots mesurée en nombre de lettres présente en effet l'avantage d'être très simple à paramétriser. C'est pourquoi elle a peu à peu pris le pas sur les approches par liste, tandis que le monde de la recherche connaissait une informatisation croissante. Les résultats présentés dans le Tableau 5.12 confirment l'efficacité de ces indices pour prédire la difficulté des textes de nos deux corpus. En effet, les meilleures corrélations de Spearman atteignent $0,525$ pour **PM8**, $0,521$ pour **NL90P** et $0,483$ pour **NLM**. À nouveau, ce sont les prédicteurs reposant sur un seuil (percentile ou nombre de lettres) qui se comportent le mieux. Par contre, une fois n'est pas coutume, la moyenne se révèle ici plus efficace que la médiane, sans que les distributions des variables ne soient spécialement plus normales que celles des autres prédicteurs.

En parlant de seuil, il est intéressant de noter le résultat des expérimentations visant à déterminer le nombre optimal de lettres au-delà duquel un mot est considéré comme difficile. Le Chapitre 3 a rapporté la remarquable convergence de plusieurs études (Barker et Stokes, 1968; Björnsson, 1983; Caylor *et al.*, 1973) sur le nombre 6, en ce qui concerne l'anglais. Pour le français, dont les mots sont en moyenne plus longs (Kandel et Moles, 1958), on s'attend à obtenir un résultat légèrement supérieur. Or, les meilleures corrélations pour les mesures **PM** sont obtenues par **PM7** et **PM8** (selon qu'on prenne en compte r ou ρ). Ce résultat apparaît donc parfaitement cohérent par rapport à la littérature.

Au-delà de l'efficacité des variantes au sein de cette famille, il est intéressant de comparer

	Test6CE				Test9CE			
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
NLM	0.479^3	$\mathbf{0.483}^3$	0.028	0.084	0.431^3	$\mathbf{0.44}^3$	0.048	0.027
NMLM	0.276^3	0.297^3	$< 0,001$	0.363	0.244^3	0.269^3	$< 0,001$	0.492
PM4	0.433^3	0.424^3	$< 0,001$	0.078	0.408^3	0.405^3	$< 0,001$	0.011
PM5	0.474^3	0.474^3	$< 0,001$	0.08	0.434^3	0.445^3	0.002	0.011
PM6	0.467^3	0.471^3	0.014	0.131	0.418^3	0.432^3	0.01	0.031
PM7	$\mathbf{0.508}^3$	0.515^3	0.015	0.053	0.458^3	0.469^3	0.002	0.005
PM8	0.504^3	$\mathbf{0.525}^3$	$< 0,001$	0.043	0.463^3	0.485^3	$< 0,001$	0.002
PM9	0.48^3	0.509^3	$< 0,001$	0.074	0.435^3	0.465^3	$< 0,001$	0.012
PM10	0.433^3	0.477^3	$< 0,001$	0.058	0.399^3	0.442^3	$< 0,001$	0.081
PM11	0.414^3	0.474^3	$< 0,001$	0.053	0.371^3	0.439^3	$< 0,001$	0.03
PM12	0.394^3	0.478^3	$< 0,001$	0.281	0.368^3	0.46^3	$< 0,001$	0.225
NL3Q	0.487^3	0.485^3	$< 0,001$	0.07	0.455^3	0.451^3	$< 0,001$	0.007
NL80P	0.491^3	0.489^3	$< 0,001$	0.075	0.453^3	0.452^3	$< 0,001$	0.026
NL90P	0.519^3	$\mathbf{0.521}^3$	$< 0,001$	0.022	0.478^3	$\mathbf{0.485}^3$	$< 0,001$	0.021

Tableau 5.12 – Mesures du voisinage orthographique et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : $^1 : < 0.05$; $^2 : < 0.01$; $^3 : < 0.001$.

celle-ci avec les autres familles lexicales. En consultant le Tableau 5.4, il est clair que les approches par liste se conduisent nettement mieux sur nos données que les variables de longueur. À ce niveau, nous nous situons donc à contre-courant des travaux des années 60 et 70, qui considéraient que la longueur moyenne des mots était préférable au nombre de termes absents. De même, les modèles unigrammes produisent des variables légèrement supérieures à celles basées sur la longueur des mots, bien que ce soit au prix d'une modélisation plus complexe.

Il a également été dit au Chapitre 3 que, pour Mesnager (1989), ce type de mesures ne serait qu'un artefact des approches par listes ou par fréquence, tandis qu'Henry (1975) défend l'existence d'un effet spécifique, lesquel semble confirmé par des études en psychologie cognitive de la lecture (O'Regan et Jacobs, 1992; Vitu *et al.*, 1990). Sans qu'il soit possible de trancher cette question à partir de nos données, les corrélations croisées de Spearman entre ces trois approches apportent des enseignements à ne pas dédaigner. Ainsi, il ressort que la force des associations entre chaque paire de ces trois facteurs, calculée sur un échantillon de *Corpus6CE*, est très similaire : $-0,65$ entre **PM8** et **ML3** ; $0,68$ entre **PM8** et **PA_Alterego1a** ; -0.66 pour **PA_Alterego1a** et **ML3**. Par comparaison, deux variables de la même famille, **PM8** et **NLM**, sont corrélées à $0,87$. On peut donc en conclure que chacune de ces trois familles de prédicteurs apporte bien une part d'information spécifique, dont l'utilité en lisibilité varie sensiblement. Il s'agit évidemment là d'une constatation essentielle pour notre propos.

Conclusion

Pour conclure l'analyse des variables lexicales, il convient d'insister sur les quelques résultats d'importance auxquels elle a donné lieu. Le plus capital d'entre eux est sans doute la supériorité de l'approche par liste dans le cadre du FLE, dont les variables, paramétrisées à l'aide d'une simple liste de mots provenant d'un manuel, surpassent l'ensemble des autres prédicteurs lexicaux. Cette observation est évidemment essentielle en ce sens qu'elle remet en question la hiérarchie des prédicteurs lexicaux plus ou moins avalisée au sein de la discipline. Il est malheureusement impossible d'affirmer que cette prédominance est intrinsèque au contexte du FLE et qu'elle ne provient pas d'un effet de corpus. Il s'agit toutefois d'une constatation très utile, d'autant qu'il s'agit d'une variable simple à paramétriser automatiquement.

À l'inverse, un autre fait notable est la faible capacité prédictive des modèles n-grammes, pourtant vantés dans la littérature. Ce résultat intrigant demanderait de plus amples explorations, afin de vérifier, d'une part, s'il ne provient pas d'une erreur d'implémentation ou d'un lissage inadéquat ; d'autre part, si l'hypothèse sur laquelle les variables de n-grammes reposent est bien valide. Cette perspective n'a cependant pas pu être explorée dans cette étude, car elle demande d'effectuer des tests sur des lecteurs.

Plusieurs autres tendances transparaissent encore de ces analyses. Tout d'abord, l'apport de connaissances spécifiques au domaine à modéliser (le FLE) – dans ce cas, une liste des mots considérés comme les plus basiques par les concepteurs d'un manuel – se traduit effectivement par des performances élevées. Par contre, l'emploi de bases lexicales plus générales et acquises à partir de documents de L1 est plus hasardeux. S'il donne lieu à des corrélations satisfaisantes pour ce qui est des listes de Gougenheim ou de Lexique3, ce n'est pas le cas des références de n-grammes. Nous avons donc tout lieu de penser qu'une liste de fréquences apprise sur des documents destinés à des apprenants conduirait à des meilleures performances. Cependant, cette perspective n'a pas pu davantage être explorée, par manque de données textuelles, cette fois.

Par ailleurs, nous avons noté qu'il est plus efficace de définir une mesure basée sur un seuil, qui considère uniquement les phénomènes dépassant ce seuil (nombre d'absents d'une liste, 75^{e} et 90^{e} percentiles des fréquences, proportion de mots de plus de 7ou 8 lettres, etc.), que d'employer les moyennes, lesquelles offrent une synthèse de la proportion d'événements faciles et difficiles dans un texte. Ces dernières semblent en effet avoir tendance à atténuer l'effet des événements d'où provient l'essentiel de la complexité. Cependant, le défaut de l'approche par seuil est que la capacité de discrimination de ces variables évolue en fonction de la position de ce seuil (et probablement du corpus d'entraînement). Réapparaît alors la

question de savoir s'il ne serait pas plus intéressant d'entraîner des formules pour chaque niveau d'apprenants, comme l'a fait Henry (1975).

Enfin, il est apparu que la grande majorité des variables ne sont pas normales. Cette observation pourrait s'expliquer par la présence de certains « outliers » moins prononcés qui auraient échappé à la procédure de suppression automatique. Toutefois, l'ampleur du phénomène nous laisse penser qu'il est plus probable que, comme l'avait souligné Bormuth (1966), nombre de prédicteurs ne suivent pas une distribution normale. De même, nos résultats confirment également ceux de ce chercheur américain au niveau de la non-linéarité de certaines variables. En effet, pas moins de la moitié d'entre elles présentent un score F significatif.

De toutes ces expériences, on donc peut retenir, en gros, que les variables lexicales comportent non seulement plusieurs prédicteurs efficaces, mais également que les diverses familles distinguées apporte effectivement une information partiellement spécifique, qui les rend complémentaires. Il s'agit évidemment là d'une constatation importante pour notre propos.

5.2.5 Analyse des variables syntaxiques

Dans cette seconde section, nous abordons la problématique des variables syntaxiques, parmi lesquelles trois familles ont été implémentées : les mesures de la longueur, les proportions de diverses catégories de discours et différentes caractéristiques flexionnelles des verbes. Pour l'exhaustivité de l'analyse, on regrettera l'absence de mesures de la complexité syntaxique, mais les raisons qui ont présidé à ce choix ont été longuement argumentées au Chapitre 3 et nous ne reviendrons pas dessus.

Comme pour les facteurs lexicaux, la pertinence de chaque prédicteur syntaxique est évaluée à l'aide des quatre mesures décrites à la section 5.2.2. L'objectif de l'analyse reste toujours de détecter les indices de la difficulté les plus efficaces en vue d'une formule de lisibilité, ainsi que de prendre position par rapport à la littérature, lorsque celle-ci existe. Rappelons en effet que les variables verbales n'ont jamais, à notre connaissance, fait l'objet d'une étude fouillée qui ait été rapportée dans la littérature.

Mesures de la longueur des unités syntaxiques

Le nombre total de mots dans un ensemble de phrases a été utilisé depuis longtemps (cf. Vogel et Washburne, 1928) pour évaluer la lisibilité de textes. Par la suite, d'autres approches de la longueur phrastique ont été développées : prendre en compte la médiane, un intervalle

entre deux percentiles (Dolch, 1948) ou la proportion de phrases plus longues qu'un seuil donné (Daoust *et al.*, 1996). Bien que critiquées théoriquement (*cf.* Chapitre 3), ces variantes ont régulièrement fait preuve d'une grande efficacité dans la littérature. C'est aussi le cas sur nos propres données, comme l'illustrent les résultats du Tableau 5.13. En effet, les deux meilleures variables, **NMP** et **NWS_902** obtiennent respectivement des ρ de $0,652$ et $0,655$ sur le *Corpus9CE*.

À nouveau, on observe une différence nette entre les coefficients de corrélation de Pearson et de Spearman, comme pour les mesures de la fréquence lexicale. Cela suggère également une relation monotone croissante non linéaire entre les variables de cette famille et la difficulté, comme le confirment les valeurs de F au test de Guilford [5] et la Figure 5.5. Sur celle-ci, on voit effectivement que l'ajustement auquel correspond le r (en rouge), ne suit pas exactement le meilleur ajustement, obtenu d'un modèle de régression polynomial local et pondéré que nous avons déjà utilisé précédemment. Nous allons donc baser notre analyse uniquement sur le ρ.

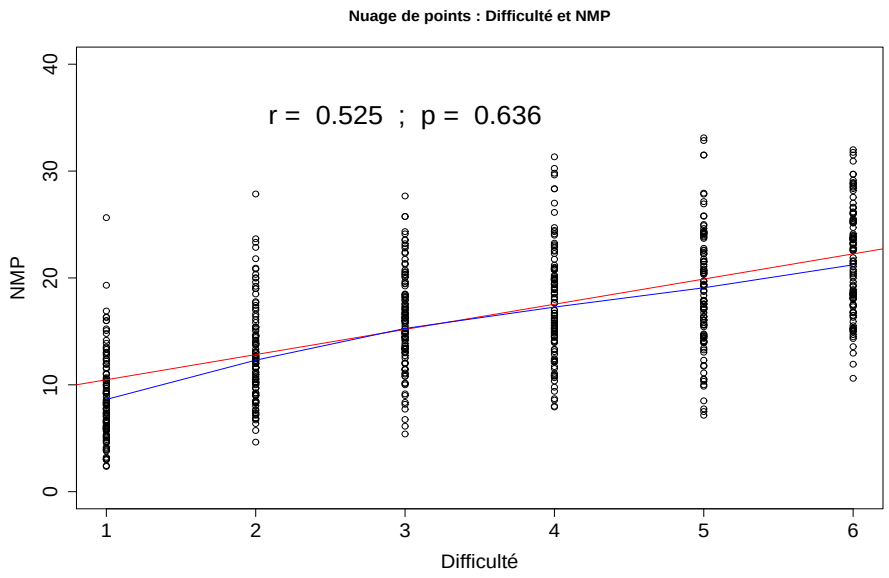


Figure 5.5 – Nuage de points entre NMP et la difficulté des textes

La première problématique à aborder touche à la définition même de la phrase. En effet, nous avions noté que le TreeTagger, employé pour distinguer les fins de phrases, présente une tendance à négliger les trois points de suspension comme le signe d'une délimitation phrastique potentielle. C'est pourquoi, à côté des indices basés sur une définition restrictive de la phrase, une version *bis* de chacun d'eux a été définie. Cette dernière segmente le texte

5. Notons que l'inverse n'est pas vrai, puisque les modèles unigrammes obtiennent un F significatif, sans que les corrélations de Spearman ne surpassent celles de Pearson.

en phrases en tenant également compte des trois points de suspension. Or, il s'avère que cette seconde façon de procéder produit des prédicteurs systématiquement plus discriminants.

Une seconde interrogation, désormais habituelle dans cette étude, concerne le choix de la meilleure mesure statistique. Cette fois, il apparaît qu'aussi bien la moyenne (**meanNWS** et **NMP**) que les percentiles en fin de distribution se comportent bien, en particulier le 90e percentile (**NWS_902** et **NWS_90**). Ils obtiennent respectivement des corrélations de $\rho = 0,618$ (**NMP**) et $\rho = 0,612$ (**NWS_902**). À côté de ces statistiques classiques, nous avons également évalué d'autres mesures moins communes, qui sont identifiées par le nom des chercheurs les ayant proposées. En ce qui concerne la proposition de Dolch (1948), qui combine une statistique de situation (médiane) et une de dispersion (90e percentile), force est de constater qu'elle obtient de moins bons scores que l'une et l'autre de ces statistiques. Quant à **LenSentDaoust**, elle est plus efficace, sans que ce soit une surprise, puisqu'elle revient à une approche par seuil, assez similaire aux percentiles.

En conclusion, il apparaît qu'aussi bien **NMP** et **NWS_902** devraient pouvoir être employées avec le même bonheur. Cette similarité des performances s'explique malheureusement par une corrélation croisée très importante. D'après les données d'un échantillon de textes issus de *Corpus6CE*, elle atteindrait $0,927$, ce qui est nettement plus élevé que dans le cas des mesures lexicales (par exemple, **MeanFFFDC** et **X75FFFDC**) ont un ρ de $0,573$. Il est donc très probable qu'un seul facteur issu de cette famille suffise pour notre formule.

	Test6CE				Test9CE			
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
meanNWS	$\mathbf{0.471}^3$	$\mathbf{0.6}^3$	$< 0,001$	0.009	$\mathbf{0.475}^3$	$\mathbf{0.635}^3$	$< 0,001$	0.015
medNWS	0.413^3	0.539^3	$< 0,001$	0.022	0.417^3	0.57^3	$< 0,001$	0.03
NWS_75	0.435^3	0.577^3	$< 0,001$	0.01	0.443^3	0.62^3	$< 0,001$	0.049
NWS_90	0.444^3	$\mathbf{0.588}^3$	$< 0,001$	0.004	0.459^3	$\mathbf{0.633}^3$	$< 0,001$	0.026
LenSentDolch	0.339^3	0.462^3	$< 0,001$	0.015	0.366^3	0.501^3	$< 0,001$	0.119
LenSentDaoust	0.456^3	0.541^3	$< 0,001$	0.214	0.467^3	0.554^3	$< 0,001$	0.528
meanNWS2 = NMP	$\mathbf{0.486}^3$	$\mathbf{0.618}^3$	$< 0,001$	0.014	$\mathbf{0.487}^3$	$\mathbf{0.652}^3$	$< 0,001$	0.031
medNWS2	0.44^3	0.545^3	$< 0,001$	0.043	0.432^3	0.571^3	$< 0,001$	0.038
NWS_752	0.471^3	0.592^3	$< 0,001$	0.004	0.466^3	0.63^3	$< 0,001$	0.066
NWS_902	$\mathbf{0.492}^3$	$\mathbf{0.612}^3$	$< 0,001$	0.003	$\mathbf{0.491}^3$	$\mathbf{0.655}^3$	$< 0,001$	0.041
LenSentDolch2	0.384^3	0.489^3	$< 0,001$	0.008	0.401^3	0.533^3	$< 0,001$	0.237
LenSentDaoust2	0.469^3	0.56^3	$< 0,001$	0.324	0.477^3	0.572^3	$< 0,001$	0.6

Tableau 5.13 – Variables de la longueur des phrases et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : $^1 : < 0.05$; $^2 : < 0.01$; $^3 : < 0.001$.

Les catégories du discours

Au sein de la famille basée sur les catégories du discours, nous avons distingué deux façons de procéder. La première, utilisée par les précurseurs en lisibilité (Gray et Leary, 1935; Vogel et Washburne, 1928), se limite à mesurer la représentation d'une catégorie donnée au sein de l'ensemble des mots du texte. La seconde, initiée par Bormuth (1966), considère l'ensemble combinatoire des ratios entre deux catégories, ce qui augmente les chances de détecter un facteur efficient. Comme la seconde démarche englobe la première, il va de soi de privilégier celle-ci. Seule ombre au tableau, le nombre de résultats devient alors très important : nous avons ainsi défini pas moins de 156 variables (*cf.* Chapitre 3.3.2). Pour éviter de surcharger inutilement cette section, seules les variables discutées dans le texte ont été reprises dans le Tableau 5.14 ; les autres sont consultables à l'annexe C. Par ailleurs, la discussion a été structurée en fonction des deux façons de procéder décrites ci-dessus.

Proportions d'une seule catégorie : En ce qui concerne les proportions simples, nous avons confronté nos résultats à ceux de trois études précédentes. Vogel et Washburne (1928) avaient ainsi comparé le pourcentage de verbes, de prépositions et de conjonctions et avaient conclu à l'infériorité de cette dernière catégorie. Bien que moins élevées, nos corrélations reproduisent ce schéma : **VER/ALL** présente un ρ de -0.193, tandis que **PRE/ALL** atteint 0.324, ce qui constitue d'ailleurs la meilleure association au sein des proportions des catégories. Toutes deux sont donc nettement supérieures au coefficient de **CON/ALL** ($\rho = 0.073$). On peut supposer que c'est parce qu'elles représentent une estimation de la complexité des structures syntaxiques, même si l'inefficacité des conjonctions est alors difficilement explicable. On s'attendrait en effet à ce que les mots introduisant des propositions coordonnées ou subordonnées soient également associés à la difficulté textuelle.

Une seconde expérience, menée par Richaudeau (1984, 89) à partir de la mémorisation de phrases, avait révélé que les adjectifs et adverbes sont plus facilement oubliés que les autres catégories du discours. Néanmoins, cet effet n'apparaît pas sur nos données : **ADJ/ALL** et **ADV/ALL** ne sont pas du tout associés avec la difficulté des textes. Peut-être cette divergence de résultats s'explique-t-elle en raison de la divergence des critères employés dans notre étude et celle de Richaudeau.

Quant aux deux supercatégories que nous avons constituées par rassemblement des mots lexicaux (CGL) ou des mots grammaticaux (CGG), nos expériences reproduisent les conclusions de Gray et Leary (1935). Ceux-ci avaient détecté une corrélation relativement basse et négative pour les mots lexicaux, laquelle se retrouve sur nos données ($r = -0,182$), et la même corrélation, mais inversée, pour les mots grammaticaux ($r = 0.182$). Les deux valeurs sont évidemment les mêmes, puisque le premier ratio est simplement l'inverse de l'autre.

Pour illustration, la figure 5.6 montre la proportion de mots grammaticaux par textes et par niveaux. On remarque que la droite d'ajustement est presque horizontale, ce qui explique le faible effet de ce facteur. D'autre part, la proportion de mots grammaticaux devient de plus en plus stable avec l'augmentation de la difficulté. Il semblerait que les textes pour natifs, que l'on rencontre en nombre dans les niveaux avancés, respectent donc un certain équilibre entre les deux catégories, ce qui rend peu discriminants les indices basés sur celles-ci.

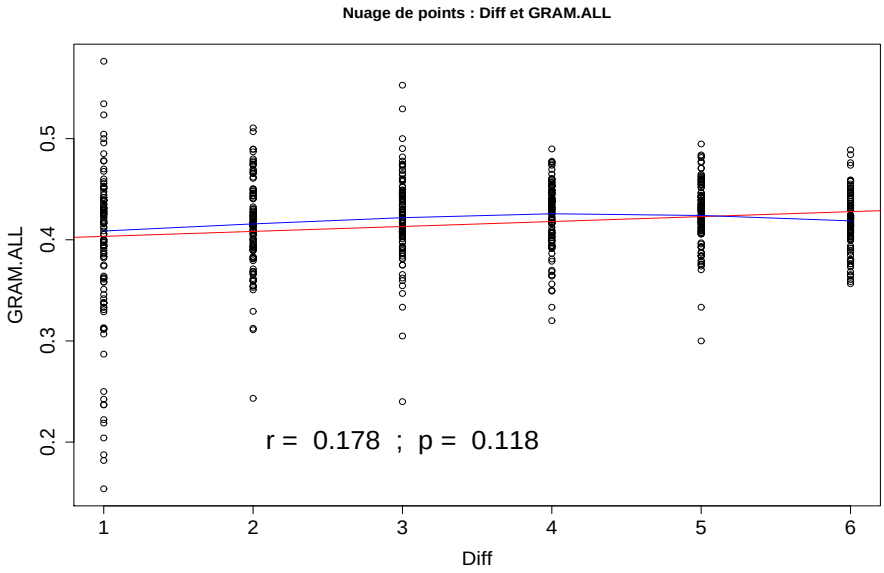


Figure 5.6 – Nuage de points entre NMP et la difficulté des textes

Les ratios entre deux catégories : L'intérêt des ratios entre deux catégories a été mis en évidence par Bormuth (1966), lequel avait trouvé que les performances de certains d'entre eux dépassaient celles des simples proportions de catégories. Nos expériences reproduisent ce constat, puisque les meilleures corrélations parmi les 156 prédicteurs sont effectivement des ratios, tous basés sur la catégorie des pronoms : **GRAM/PRO** ($\rho = 0,342$) ; **PRE/PRO** ($\rho = 0,345$) et **ART/PRO** ($\rho = 0,326$). Cette observation est particulièrement digne d'intérêt si l'on se souvient que le pourcentage de pronoms personnels a depuis longtemps été considéré comme un prédicteur fiable en soi et notamment associé au niveau de personnalisation du texte. Il semblerait que, plus généralement, les pronoms constituent la catégorie de mots la plus liée à la difficulté des textes. La corrélation de Pearson de **PRO/ALL** ($r = -0,347$) est d'ailleurs la plus élevée parmi les proportions d'une catégorie donnée dans les textes.

De manière plus générale, les ratios qui sortent du lot font tous intervenir au moins une catégorie classée parmi les mots grammaticaux, telle que les pronoms, les articles ou les prépositions. Ainsi, **PRE/LEX** atteint un ρ de $0,322$; **PRE/GRAM**, $0,304$ et **ART/LEX**,

$0,305$. Seules exceptions, les ratios basés sur les conjonctions obtiennent tous des corrélations très faibles, voire inexistantes. Ces chiffres révèlent aussi l'intérêt des ratios basés sur les supercatégories plutôt que sur une seule catégorie. La catégorie issue de la fusion des noms communs et des noms propres (NOM.NAM) présente également un certain pouvoir discriminant lorsqu'elle est combinée avec l'article (ρ de **NOM.NAM/ART** = $-0,29$) ou la préposition (ρ de **NOM.NAM.PRE** $= -0,144$).

	Test6CE				Test9CE			
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
ADJ.ALL	0.026	0.102^{1}	$< 0,001$	0.146	0.013	0.085	$< 0,001$	0.135
ART.ALL	0.309^{3}	0.308^{3}	$< 0,001$	0.132	0.28^{3}	0.288^{3}	$< 0,001$	0.42
CON.ALL	0.074	0.073	$< 0,001$	0.293	0.127^{3}	0.132^{3}	$< 0,001$	0.068
NAM.ALL	-0.23^{3}	-0.147^{2}	$< 0,001$	0.007	-0.282^{3}	-0.202^{3}	$< 0,001$	$< 0,001$
NOM.ALL	0.213^{3}	0.219^{3}	$< 0,001$	0.435	0.212^{3}	0.221^{3}	$< 0,001$	0.595
PRE.ALL	$\mathbf{0.341}^{3}$	$\mathbf{0.324}^{3}$	$< 0,001$	0.192	$\mathbf{0.364}^{3}$	$\mathbf{0.353}^{3}$	$< 0,001$	0.009
PRO.ALL	$\mathbf{-0.347}^{3}$	-0.313^{3}	$< 0,001$	0.136	-0.328^{3}	-0.305^{3}	$< 0,001$	0.161
VER.ALL	-0.199^{3}	-0.193^{3}	$< 0,001$	0.066	-0.159^{3}	-0.16^{3}	$< 0,001$	0.067
GRAM.ALL	0.182^{2}	0.128^{1}	$< 0,001$	0.025	0.221^{3}	0.178^{3}	$< 0,001$	0.005
LEX.ALL	-0.182^{2}	-0.128^{1}	$< 0,001$	0.025	-0.221^{3}	-0.178^{3}	$< 0,001$	0.005
ADJ.NAM	0.012	0.171^{3}	$< 0,001$	0.52	0.034	0.211^{3}	$< 0,001$	0.586
ADJ.PRO	-0.065^{1}	0.272^{3}	$< 0,001$	0.302	-0.06	0.25^{3}	$< 0,001$	0.649
ADJ.VER	-0.02	0.163^{3}	$< 0,001$	0.264	-0.024	0.141^{2}	$< 0,001$	0.457
ADV.PRO	0.006	0.312^{3}	$< 0,001$	0.503	-0.009	0.284^{3}	$< 0,001$	0.704
ART.LEX	0.305^{3}	0.311^{3}	$< 0,001$	0.128	0.283^{3}	0.298^{3}	0.007	0.411
ART.PRO	-0.02	$\mathbf{0.326}^{3}$	$< 0,001$	0.319	-0.024	0.318^{3}	$< 0,001$	0.503
CON.PRO	-0.02	0.319^{3}	$< 0,001$	0.512	-0.011	$\mathbf{0.335}^{3}$	$< 0,001$	0.578
GRAM.PRE	-0.171^{3}	-0.304^{3}	$< 0,001$	0.269	-0.192^{3}	-0.316^{3}	$< 0,001$	0.023
GRAM.PRO	-0.028	$\mathbf{0.342}^{3}$	$< 0,001$	0.344	-0.03	$\mathbf{0.343}^{3}$	$< 0,001$	0.574
LEX.ART	-0.158^{3}	-0.311^{3}	$< 0,001$	0.07	-0.158^{3}	-0.298^{3}	$< 0,001$	0.01
LEX.PRE	-0.158^{3}	$\mathbf{-0.322}^{3}$	$< 0,001$	0.19	-0.183^{3}	$\mathbf{-0.36}^{3}$	$< 0,001$	0.002
NOM.NAM.ART	-0.143^{2}	-0.29^{3}	$< 0,001$	0.067	-0.148^{3}	-0.279^{3}	$< 0,001$	0.004
NOM.NAM.PRE	-0.144^{2}	-0.289^{3}	$< 0,001$	0.181	-0.166^{3}	-0.345^{3}	$< 0,001$	$< 0,001$
NOM.PRO	-0.05	0.305^{3}	$< 0,001$	0.371	-0.054	0.297^{3}	$< 0,001$	0.56
PRE.PRO	-0.034	$\mathbf{0.345}^{3}$	$< 0,001$	0.35	-0.038	$\mathbf{0.349}^{3}$	$< 0,001$	0.597
PRE.GRAM	0.282^{3}	0.304^{3}	$< 0,001$	0.324	0.301^{3}	0.316^{3}	$< 0,001$	0.035
PRE.LEX	0.342^{3}	0.322^{3}	$< 0,001$	0.16	0.367^{3}	0.36^{3}	$< 0,001$	0.009
PRO.ART	-0.165^{3}	$\mathbf{-0.326}^{3}$	$< 0,001$	0.136	-0.153^{3}	$\mathbf{-0.317}^{3}$	$< 0,001$	0.05
PRO.GRAM	-0.389^{3}	$\mathbf{-0.342}^{3}$	$< 0,001$	0.089	-0.382^{3}	-0.343^{3}	$< 0,001$	0.012
PRO.LEX	-0.316^{3}	-0.286^{3}	$< 0,001$	0.17	-0.29^{3}	-0.273^{3}	$< 0,001$	0.199
PRO.NOM	-0.276^{3}	-0.305^{3}	$< 0,001$	0.284	-0.223^{3}	-0.297^{3}	$< 0,001$	0.305
PRO.PRE	-0.181^{3}	$\mathbf{-0.345}^{3}$	$< 0,001$	0.226	-0.194^{3}	$\mathbf{-0.349}^{3}$	$< 0,001$	0.021
VER.ART	-0.172^{3}	-0.289^{3}	$< 0,001$	0.179	-0.16^{3}	-0.266^{3}	$< 0,001$	0.033
VER.NOM	-0.219^{2}	-0.226^{3}	$< 0,001$	0.364	-0.173^{2}	-0.209^{3}	$< 0,001$	0.71
VER.PRE	-0.195^{3}	-0.313^{3}	$< 0,001$	0.268	-0.19^{3}	-0.316^{3}	$< 0,001$	0.021

Tableau 5.14 – Variables basées sur les catégories grammaticales et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : ^{1} : < 0.05 ; ^{2} : < 0.01 ; ^{3} : < 0.001.

Soulignons enfin le ratio d'adjectifs par rapport aux noms propres (**ADJ/NAM**) dont le co-

efficient n'est pas terrible ($\rho = 0,171$), mais significatif. Celui-ci donne sans doute une idée de nombre de modifieurs adjectivaux par tête nominale, ce qui se rapproche de l'hypothèse de Bloomer (1959). Celui-ci supposait que les adjectifs et adverbes servent à préciser le sens des mots vagues et que leur nombre augmenterait proportionnellement au niveau d'abstraction. Il faisait aussi référence à un article de Broder (1940, cité par Bloomer, 1959, 269) où celui-ci a détecté que le ratio **ADJ/VER** augmente également en fonction du taux d'abstraction. Celui-ci obtient sur nos données un $\rho = 0,171$. Il s'agit à nouveau d'une association tout à fait significative, mais dont la taille ne suffit pas à en faire un bon prédicteur.

C'est d'ailleurs ce qui ressort de la plupart des variables de cette famille : la majorité de celles reprises dans le Tableau 5.14 présente une corrélation qui est tout à fait significative ($p < 0,0001$), sans être particulièrement élevée. Cette caractéristique s'explique essentiellement par la quantité de textes : plus celle-ci augmente, plus il est aisé de détecter une corrélation significative entre un facteur et la difficulté. Cependant, comme nous l'avons déjà expliqué, l'existence d'une association ne suffit pas à faire d'une variable un bon prédicteur. Il faut qu'elle soit suffisamment importante et non corrélée avec un autre facteur.

À ce propos, un dernier aspect intéressant à explorer concerne l'interprétation des variables de cette famille, située à l'intersection des dimensions grammaticale et lexicale. La technique envisagée pour ce faire est simplement de comparer quelques-unes parmi celles ayant fait l'objet de la discussion (i.e. **ADJ/NOM**, **PRE/ALL**, **VER/ALL**, **PRE/PRO**, **ADJ/-NOM**) avec les meilleures variables lexicales et syntaxiques détectées jusqu'ici (*cf.* Tableau 5.15). Nous avons pour ce faire utilisé plutôt Spearman que Pearson, ayant noté que certaines corrélations croisées varient considérablement : **PA_Alterego1a** passe ainsi de $r = 0,04$ à $\rho = 0,49$.

	ADJ/NOM	PRE/ALL	PRE/GRAM	PRE/PRO	VER/ALL
ADJ/NOM	1	-0.07	0.058	0.008	-0.09
PRE/ALL	-0.07	1	0.93	0.77	-0.51
PRE/GRAM	0.06	**0.93**	1	0.83	-0.49
PRE/PRO	0,008	0.77	0.83	1	-0.67
VER/ALL	-0.09	$\mathbf{-0.50}$	$\mathbf{-0.49}$	-0.67	1
ML3	-0.01	-0.21	-0.27	-0.37	0.22
NLM	0.05	0.39	0.42	0.55	-0.43
NMP	-0.03	0.51	0.45	0.52	-0.39
PA_Alterego1a	-0.01	$\mathbf{0,42}$	0.42	0.485	-0.35

Tableau 5.15 – Corrélations croisées (ρ de Spearman) entre quelques variables catégorielles et les principaux indices lexicaux et syntaxiques, calculées sur l'un des 10 échantillons issus de Corpus6CE.

À la lecture de cette table, il apparaît que l'interprétation qui peut être donnée des ratios varie de l'un à l'autre. Ainsi, **ADJ/NOM** n'est associé ni aux indices lexicaux, ni aux syn-

taxiques. À l'inverse, **PRE/PRO** et **VER/ALL** et **PRE/ALL** obtiennent des scores similaires pour l'une et l'autre dimensions, tandis que **PRE/ALL** ressortit davantage aux aspects syntaxiques. Face à cette diversité des résultats, il n'est pas possible de conclure à une interprétation particulière des ratios. Au contraire, ils confirment à nouveau la multiplicité des aspects pris en compte par cette famille de variables.

Les aspects verbaux

La dernière famille de variables syntaxiques traite de l'apport des temps et modes verbaux à la lisibilité des textes. Les résultats, présentés dans le Tableau 5.16, mettent en évidence la capacité de discrimination de certaines caractéristiques verbales. Ainsi, la présence d'au moins un participe présent ($\rho = 0,44$), d'un participe passé ($\rho = 0,388$), voire d'un infinitif ($\rho = 0,341$) se révèle une information digne d'intérêt. De même, la proportion de participes présents ($\rho = 0,41$) ou simplement d'indicatifs présents ($\rho = -0,337$) constituent des prédicteurs potentiellement intéressants pour notre formule de lisibilité pour le FLE. Ce constat revêt une importance particulière étant donnée sa nouveauté au sein de la discipline, comme il l'a été dit au Chapitre 3.3.2.

En vue d'affiner ces premières observations, il convient de comparer les deux paramétrisations des caractéristiques verbales que nous avons implémentées. La première se base sur l'hypothèse que l'apprentissage de la conjugaison est progressif et que, par conséquent, il existe un lien entre la présence d'un temps ou d'un mode dans un texte et le niveau de difficulté auquel un apprenant risque de rencontrer ce texte. La seconde considère par contre que c'est le décodage de chaque forme qui importe et qu'un texte comportant davantage de formes conjuguées ardues est dès lors plus complexe. Il apparaît, à la lecture du Tableau 5.16, que les coefficients obtenus par l'approche binaire surpassent systématiquement ceux de l'approche continue, si ce n'est pour l'indicatif présent. Rien n'exclut néanmoins que cette supériorité ne soit un artefact du critère utilisé, puisque les textes de notre corpus proviennent de manuels de FLE, justement conçus en fonction de cette acquisition progressive. Il faudrait vérifier que cet avantage de l'approche binaire se généralise à des textes provenant d'autres sources.

En ce qui concerne le cas particulier du présent de l'indicatif, l'analyse des corrélations permet d'en pousser plus loin la compréhension. On remarque en effet que toutes les variables binaires présentent soit une corrélation positive avec la difficulté, soit une association non significative. Cela signifie que soit une caractéristique est présente dans la quasi-totalité des textes, et n'a donc pas de pouvoir prédictif, soit elle n'apparaît que dans les niveaux supérieurs. Le présent appartient clairement à la première catégorie : il représente donc un temps commun pour lequel l'information binaire ne suffit pas. Il est dès lors plus efficace de tenir

	Test6CE				Test9CE			
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
Cond	0.227^{3}	0.227^{3}	$< 0,001$	$< 0,001$	0.25^{3}	0.249^{3}	$< 0,001$	0.02
Futur	0.253^{3}	0.252^{3}	$< 0,001$	$< 0,001$	0.272^{3}	0.271^{3}	$< 0,001$	0.002
Imperatif	0.019	0.019	$< 0,001$	0.391	0.022	0.022	$< 0,001$	0.671
Impf	0.272^{3}	0.272^{3}	$< 0,001$	$< 0,001$	0.317^{3}	0.317^{3}	$< 0,001$	$< 0,001$
Infi	$\mathbf{0.342}^{3}$	$\mathbf{0.341}^{3}$	$< 0,001$	$< 0,001$	$\mathbf{0.375}^{3}$	$\mathbf{0.374}^{3}$	$< 0,001$	$< 0,001$
PPasse	$\mathbf{0.388}^{3}$	$\mathbf{0.388}^{3}$	$< 0,001$	$< 0,001$	$\mathbf{0.425}^{3}$	$\mathbf{0.424}^{3}$	$< 0,001$	$< 0,001$
PPres	$\mathbf{0.44}^{3}$	$\mathbf{0.44}^{3}$	$< 0,001$	0.003	$\mathbf{0.463}^{3}$	$\mathbf{0.463}^{3}$	$< 0,001$	0.023
Pres	0.078	0.078	$< 0,001$	0.406	0.059	0.059	$< 0,001$	0.69
PasseSim	0.146^{2}	0.146^{2}	$< 0,001$	0.017	0.155^{3}	0.155^{3}	$< 0,001$	0.125
Subi	0.049	0.049	$< 0,001$	0.608	0.076^{1}	0.076^{1}	$< 0,001$	0.575
Subp	0.266^{3}	$\mathbf{0.266}^{3}$	$< 0,001$	0.423	0.247^{3}	0.247^{3}	$< 0,001$	0.438
Cond_C	0.023	0.192^{3}	$< 0,001$	0.197	0.043	0.218^{3}	$< 0,001$	0.252
Futur_C	0.065	0.219^{3}	$< 0,001$	0.086	0.064	0.239^{3}	$< 0,001$	0.021
Imperatif_C	-0.022	0.019	$< 0,001$	0.413	-0.032	0.021	$< 0,001$	0.681
Impf_C	0.024	0.223^{3}	$< 0,001$	0.076	0.088^{1}	0.275^{3}	$< 0,001$	$< 0,001$
Infi_C	0.232^{3}	0.282^{3}	$< 0,001$	0.069	0.273^{3}	0.331^{3}	$< 0,001$	0.006
PPasse_C	0.262^{3}	0.304^{3}	$< 0,001$	0.003	0.31^{3}	0.366^{3}	$< 0,001$	$< 0,001$
PPres_C	0.252^{3}	$\mathbf{0.41}^{3}$	$< 0,001$	0.123	0.246^{3}	$\mathbf{0.432}^{3}$	$< 0,001$	0.205
Pres_C	$\mathbf{-0.355}^{3}$	$\mathbf{-0.337}^{3}$	$< 0,001$	$< 0,001$	$\mathbf{-0.439}^{3}$	$\mathbf{-0.433}^{3}$	$< 0,001$	$< 0,001$
PasseSim_C	-0.012	0.121^{1}	$< 0,001$	0.086	-0.002	0.13^{3}	$< 0,001$	0.039
Subi_C	-0.023	0.047	$< 0,001$	0.494	0.007	0.074^{1}	$< 0,001$	0.409
Subp_C	-0.07	0.091	$< 0,001$	0.263	-0.073^{1}	0.09^{1}	$< 0,001$	0.301
ratioDaou	-0.288^{3}	-0.287^{3}	$< 0,001$	0.193	-0.258^{3}	-0.264^{3}	$< 0,001$	0.22

Tableau 5.16 – Caractéristiques verbales et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : ^{1} : < 0.05 ; ^{2} : < 0.01 ; ^{3} : < 0.001.

compte de la proportion de verbes conjugués au présent de l'indicatif, car la présence d'un grand nombre d'entre eux dans un texte rend celui-ci plus accessible, comme le confirme le signe négatif de la corrélation entre **Pres** et la difficulté. Cette différence de comportement de **Pres** entre l'approche binaire et continue constitue l'un des résultats les plus intéressants de cette famille de prédicteurs.

Cependant, au-delà de la découverte que les informations de temps verbaux sont tout à fait pertinentes en lisibilité du FLE, il convient d'analyser leurs rapports avec des prédicteurs plus classiques, afin de vérifier qu'il ne s'agit pas d'un nouveau « proxy » pour une dimension linguistique déjà amplement mesurée par ailleurs. À nouveau, les corrélations croisées des principaux indices de cette famille ont été mises en correspondance avec trois des meilleures variables, ainsi qu'avec eux-mêmes (*cf.* Tableau 5.17). Il en ressort que si les temps du présent (indicatif et participe) sont moyennement corrélés aussi bien avec les variables lexicales que syntaxiques, d'autres (**Subp** ou **Infi**) entretiennent avec elles des associations faibles. Ce recoupement informationnel limité laisse présager une complémentarité intéressante entre ces différents facteurs, qui pourrait permettre de dépasser les performances des formules classiques.

	Infi	PPasse	PPres	Subp	Pres_C
ML3	$-0,148$	$-0,23$	$\mathbf{-0,37}$	$-0,08$	$\mathbf{0,35}$
NMP	$0,27$	$\mathbf{0,33}$	$\mathbf{0,43}$	$0,18$	$\mathbf{-0,42}$
PA_Alterego1a	$0,24$	0.27	$\mathbf{0.36}$	$0,13$	$\mathbf{-0,32}$
Infi	1	$0,28$	$0,17$	$0,17$	$-0,24$
PPasse	$0,28$	1	$0,18$	$0,15$	$-0,23$
PPres	$0,17$	$0,18$	1	$0,17$	$-0,36$
Subp	$0,17$	$0,15$	$0,17$	1	$-0,01$
Pres_C	$-0,24$	$-0,23$	$\mathbf{-0,36}$	$-0,01$	1

Tableau 5.17 – Corrélations croisées (ρ de Spearman) entre les meilleures variables verbales et trois indices lexicaux et syntaxiques typiques, calculées sur l'un des 10 échantillons issus de Corpus6CE, choisi au hasard.

Terminons la discussion des aspects verbaux par la variable proposée par Gillie (1957) et Daoust *et al.* (1996), à savoir le rapport entre le nombre de verbes conjugués sur le nombre de mots du texte. Il est intéressant de noter que cette variable plus classique est moins efficace que nos propres indices : son ρ ne dépasse pas -0.287. Le signe négatif indique également que le nombre de verbes conjugués diminue avec l'augmentation de la difficulté, ce qui n'est pas particulièrement surprenant. En effet, les textes les plus simples comportent surtout des phrases simples dont les verbes sont conjugués au présent, tandis que ceux pour apprenants avancés comprennent davantage de propositions infinitives.

Conclusion

Pour en conclure avec les indices de la difficulté syntaxique, rappelons que plusieurs variables intéressantes ont été détectées au sein de chacune des trois familles. Ces bons candidats font surtout intervenir les pronoms (**PRE/PRO** et **GRAM/PRO**), les participes (**PPasse**, **PPres**, **PPres_C**) et le nombre de mots par phrase (**NMP** et **NWS_902**). Un atout supplémentaire de ces facteurs provient du fait que les corrélations qu'ils entretiennent, que ce soit au sein d'une même famille ou entre différentes familles, restent de taille moyenne ou faible. Cela devrait donc permettre d'en conserver plusieurs au sein d'un même modèle.

Par ailleurs, il faut remarquer que la supériorité des variables basées sur des seuils et sur des mesures de dispersion, détectée pour le niveau lexical, n'a pas été reproduite pour la dimension syntaxique. Non seulement **NMP** se révèle aussi prédictif que **NWS_902**, mais ces deux indices sont davantage intercorrélés que ne le sont les mesures lexicales de tendance centrale et de dispersion. Quant aux variables catégorielles, l'analyse de leurs corrélations croisées a mis en évidence la multiplicité de leurs interprétations, ce qui avait déjà été rapporté par la littérature. Il semblerait qu'il en aille de même en ce qui concerne les aspects verbaux. Dans les deux cas, une étude plus poussée de cette question pourrait néanmoins se révéler instructive.

5.2.6 Analyse des variables sémantiques

Cette troisième section expose l'ensemble des variables sémantiques implémentées dans cette étude, qui correspondent tantôt au taux de personnalisation, tantôt à la densité des idées, tantôt encore à la cohérence interne aux textes à évaluer. Bien qu'elle ne soit pas aussi étoffée que les deux sections précédentes, pour les raisons expliquées au Chapitre 3, cette partie présente par contre un contenu plus innovant. En effet, il n'y a pas, à notre connaissance, d'études sur le français et *a fortiori* sur le FLE qui investiguent de manière approfondie les prédicteurs du niveau sémantique. Pas plus qu'il n'existe, pour cette langue, de formule qui les intègre.

Or, nous avons insisté précédemment (*cf.* Chapitre 2.4) sur notre désir d'inclure cette étude dans le paradigme de la lisibilité computationnelle, lequel intègre des technologies de TAL et vise à les employer afin, notamment, de mieux modéliser les aspects sémantiques. Ce faisant, il pourrait constituer une synthèse des approches classique et structuro-cognitive. Toutefois, plusieurs résultats présentés dans le Chapitre 3 ont semé le doute sur l'efficacité des variables sémantiques lorsqu'elles sont paramétrisées automatiquement, et, par conséquent, sur la possibilité de cette synthèse. En effet, bien que plusieurs chercheurs se soient enthousiasmés pour ce type de prédicteurs, les résultats rencontrés dans la littérature nous ont semblé souvent moins optimistes que la description qui en était faite. C'est pourquoi, au-delà de la simple évaluation des prédicteurs, les analyses de cette section importent également par ce qu'elles peuvent apporter vis-à-vis de cette interrogation.

Taux d'abstraction et de personnalisation du texte

Parmi les différents indices du taux d'abstraction d'un texte qui ont été identifiés dans la littérature, deux approches ont été retenues. La première considère la proportion de certaines catégories de discours, en particulier les adjectifs, comme un moyen d'approximer le niveau d'abstraction d'un texte. Elle a déjà fait l'objet d'une discussion dans la section sur les variables catégorielles (*cf.* 5.2.5). La seconde, envisagée ici, se base sur la proportion de pronoms personnels dans le texte, laquelle serait significative de son taux de personnalisation. Les résultats sont présentés dans le Tableau 5.18.

Dans le volet théorique consacré à cette famille (*cf.* Chapitre 3.2.4), une série de résultats relativement contradictoires a été présentée. Les contradictions portaient en particulier sur le regroupement optimal des pronoms personnels au sein de différentes variables et sur le sens de l'effet que celles-ci ont sur la difficulté des textes. Par ailleurs, il n'a pas toujours été possible de déterminer précisément le type de normalisation opéré dans les études précédentes. C'est pourquoi nous avons expérimenté, pour chaque facteur, les deux normalisations pos-

	Test6CE				Test9CE			
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
PP1	-0.329^{3}	-0.224^{3}	$< 0,001$	0.044	-0.324^{3}	-0.235^{3}	$< 0,001$	0.012
PropP1	-0.153^{3}	-0.135^{2}	$< 0,001$	0.096	-0.162^{3}	-0.148^{3}	$< 0,001$	0.117
PP2	$\mathbf{-0.338}^{3}$	$\mathbf{-0.325}^{3}$	$< 0,001$	0.095	$\mathbf{-0.333}^{3}$	$\mathbf{-0.325}^{3}$	$< 0,001$	0.058
PropP2	-0.306^{3}	-0.306^{3}	$< 0,001$	0.347	-0.299^{3}	-0.303^{3}	$< 0,001$	0.251
PP3	-0.179^{3}	-0.063	$< 0,001$	0.411	-0.165^{3}	-0.042	$< 0,001$	0.57
PropP3	0.315^{3}	0.302^{3}	$< 0,001$	0.319	0.308^{3}	0.297^{3}	$< 0,001$	0.147
PP2_S	-0.229^{3}	-0.21^{3}	$< 0,001$	0.138	-0.22^{3}	-0.213^{3}	$< 0,001$	0.121
PropP2_S	-0.221^{3}	-0.209^{3}	$< 0,001$	0.21	-0.208^{3}	-0.212^{3}	$< 0,001$	0.138
PP2_P	-0.249^{3}	-0.246^{3}	$< 0,001$	0.233	-0.248^{3}	-0.242^{3}	$< 0,001$	0.17
PropP2_P	-0.223^{3}	-0.232^{3}	$< 0,001$	0.311	-0.217^{3}	-0.225^{3}	$< 0,001$	0.139
PP1P2	$\mathbf{-0.408}^{3}$	$\mathbf{-0.333}^{3}$	$< 0,001$	0.008	$\mathbf{-0.405}^{3}$	$\mathbf{-0.346}^{3}$	$< 0,001$	$< 0,001$
PropP1P2	-0.305^{3}	-0.282^{3}	$< 0,001$	0.069	-0.315^{3}	-0.298^{3}	$< 0,001$	0.009
ADJ/VER	-0.02	0.163^{3}	$< 0,001$	0.264	-0.024	0.141^{2}	$< 0,001$	0.457

Tableau 5.18 – Caractéristiques verbales et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : ^{1} : < 0.05 ; ^{2} : < 0.01 ; ^{3} : < 0.001.

sibles : en fonction du nombre total de mots dans le texte et du nombre total de pronoms personnels. Il s'avère que la première produit systématiquement des coefficients plus élevés que la seconde, sauf pour **PropP3**. La discussion qui suit est dès lors essentiellement basée sur la première.

Cherchant à démêler les différentes contradictions précitées, nous avons tout d'abord distingué quatre classes de pronoms personnels : **PP1**, **PP2_S**, **PP2_P** et **PP3**. Il en ressort un tableau assez net qui ne s'accorde pas avec certaines observations de la littérature, mais en confirme d'autres. Tout d'abord, alors que Daoust *et al.* (1996) recommandent de distinguer les pronoms de la S2 et de la P2, car leurs effets seraient opposés (le premier faciliterait la lecture, tandis que le second la compliquerait) nos résultats ne vont pas dans ce sens. Non seulement **PP2_S** et **PP2_P** présentent tous deux un coefficient de corrélation négatif et significatif (r [6] vaut respectivement $-0,229$ et $-0,249$), mais lorsqu'on les combine en une variable unique (**PP2**), il apparaît que la quantité d'information qu'ils apportent s'accroît. Le r de **PP2** atteint alors $-0,338$, ce qui en fait la seconde meilleure variable de cette famille, après **PP1P2**.

Par conséquent, **PP2** est également supérieur aux variables basées sur les deux autres personnes : **PP1** et **PP3**. Ce motif reproduit les résultats de Dale et Tyler (1934) en ce qui concerne l'efficacité prédictive des trois classes. Par contre, notre étude se différencie de ces deux auteurs en ce qui concerne le sens des corrélations. Ceux-ci ont observé une opposition entre **PP2**, qui faciliterait la lecture, et les deux autres classes, dont la présence rendrait les textes plus complexes. De notre côté, les trois classes produisent un effet similaire : plus il

6. Pour cette famille, c'est l'usage du r de Pearson qui est employé, étant donné que la grande majorité des tests F de Guilford permettent de conclure à une relation linéaire entre ces variables et la difficulté.

y a de pronoms, plus le niveau des textes est bas. D'ailleurs, on observe à nouveau que la combinaison de **PP1** et **PP2**, à savoir **PP1P2**, dépasse les scores individuels à un point tel qu'elle en est aussi le meilleur indice de cette famille. Par là, nos résultats confirment les travaux d'Henry (1975), qui a déjà combiné ces deux classes avec succès, mais aussi ceux de Gray et Leary (1935), lesquels ont observé un effet facilitateur pour les trois classes.

Ce tableau apparaît cohérent par rapport à la littérature et permet de conclure à l'efficacité de quelques prédicteurs basés sur les pronoms personnels, en particulier **PP1P2**. Ce constat est à mettre en rapport avec la discussion portant sur les variables catégorielles, laquelle a mis en évidence l'intérêt des pronoms dans leur ensemble (**PRO/ALL**). Dès lors, il s'agit de vérifier dans quelle mesure l'information apportée par les pronoms personnels est différente de **PRO/ALL**. À nouveau, les corrélations croisées entre **PRO/ALL** et les meilleurs facteurs de cette famille ont été calculées ; elles sont rapportées dans le Tableau 5.19.

	PP1	PP2	PropP3	PP1P2	PRO/ALL
PP1	1	$0,44$	$-0,58$	$0,89$	$0,67$
PP2	$0,44$	1	$-0,48$	$0,71$	$0,56$
PropP3	$-0,58$	$-0,48$	1	$-0,67$	$-0,27$
PP1P2	$0,89$	$0,71$	$-0,67$	1	$0,72$

*Tableau 5.19 – Corrélations croisées (ρ de Spearman) entre quelques variables du taux de personnalisation textuel et **PRO/ALL**, calculées sur l'un des 10 échantillons issus de Corpus6CE.*

Il ressort de cette analyse que **PRO/ALL** est nettement corrélé avec les variables qui utilisent uniquement les pronoms personnels. Le coefficient augmente d'ailleurs logiquement lorsqu'on combine les pronoms personnels de la 1re et de la 2e personne, puisque davantage de pronoms sont alors pris en compte. Le type d'information véhiculé par ces différents prédicteurs semble donc partiellement similaire, mais aussi partiellement différent. On remarque également que, vu l'association élevée entre **PropP3** et **PP1**, le choix d'une normalisation plutôt qu'une autre ne semble pas impacter la nature de l'information. D'ailleurs, le ρ entre **PP1** et **PropP1** atteint $0,95$, tandis que **PropP2** et **PP2** culmine à $0,99$, ce qui confirme cette hypothèse. Il faut cependant noter le cas étonnant de **PropP3**, qui se distingue sensiblement de **PP3**.

La densité des idées

Le second aspect sémantique pris en compte par cette étude est la densité des idées, modélisée d'après l'analyse prédicative de Kintsch et van Dijk (1978) à l'aide du logiciel *Densidées*. Comme le Chapitre 3.2.4 l'a mis en évidence, l'efficacité de cette variable n'a jamais été véritablement attestée. Kintsch et Vipond (1979) l'insèrent dans leur formule cognitive,

sans rapporter de chiffres précis concernant son efficacité. Par la suite, Kemper *et al.* (1993) l'évaluent dans un cadre expérimental plus réaliste et n'obtiennent qu'une corrélation de $-0,27$, qui indique un faible effet négatif de la densité conceptuelle sur la difficulté des textes.

	Test6CE				Test9CE			
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
ConcDens	0.25^3	0.253^3	$< 0,001$	0.093	0.282^3	0.277^3	$< 0,001$	$< 0,001$

Tableau 5.20 – Densité des idées et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : $^1 : < 0.05$; $^2 : < 0.01$; $^3 : < 0.001$.

Les résultats du Tableau 5.20 sont étonnamment cohérents avec les observations de ces auteurs. En effet, l'association de **ConcDens** avec la difficulté sur le *Corpus6CE* ne dépasse pas $0,25$ [7], ce qui se situe probablement dans le même ordre de grandeur que le coefficient de Kemper *et al.* (1993). On peut donc conclure à un faible pouvoir prédictif de cette variable, quand bien même la corrélation est nettement significative. Par ailleurs, on note que les corrélations croisées de **ConcDens** avec les meilleures variables lexicales et syntaxiques sont faibles (*cf.* Tableau 5.21), et qu'elle est à peine supérieure avec la variable sémantique **ADJ/VER**. La densité des idées apporte donc bien une information très particulière, qui, malheureusement, ne semble guère utile pour la prédiction de la difficulté.

	ML3	NMP	PP1P2	ADJ/VER
ConcDens	$-0,24$	$0,23$	$-0,12$	30

*Tableau 5.21 – Corrélations croisées (ρ de Spearman) entre **ConcDens** et quelques variables typiques, calculées sur l'un des 10 échantillons issus de Corpus6CE.*

Le taux de cohérence mesuré via LSA

La dernière famille de variables sémantiques utilise elle aussi des technologies de TAL assez avancées en comparaison de la plupart des autres indices envisagés. Elle repose sur un espace vectoriel, transformé soit à l'aide du *tf-idf*, soit par LSA, et vise à évaluer automatiquement le niveau de cohérence interne d'un texte. L'hypothèse sous-jacente est qu'un texte plus cohérent serait plus aisé à lire.

Avant de poser un diagnostic sur l'efficacité de cette famille de variables, il a été néces-

7. Bien que le signe de la corrélation ne soit pas le même que chez Kemper *et al.* (1993), l'effet va bien en réalité dans le même sens, étant donné que cet auteur prend comme critère le taux de compréhension, qui décroît proportionnellement à l'augmentation de la difficulté.

saire de résoudre un problème d'optimisation : déterminer le nombre idéal de dimensions à conserver dans l'espace modifié de la LSA. Comme nous l'avons déjà expliqué, le choix de ce nombre, exprimé au moyen du paramètre k (*cf.* Chapitre 3.3.3), est crucial pour la LSA dont dépendent les performances. Il a donc été nécessaire de mettre en place une méthodologie permettant de sélectionner un k relativement optimal, sachant qu'au vu du temps nécessaire pour entraîner un modèle par LSA, il n'est pas possible d'explorer l'ensemble des valeurs possibles.

La stratégie utilisée pour déterminer le meilleur k est une « grid search », guidée manuellement et opérée à deux niveaux de granularité sur le *Corpus6CE*. Elle consiste à choisir plusieurs valeurs initiales parmi le spectre des possibles (c'est-à-dire 1 et 1433, le nombre de documents), selon des intervalles assez larges. Le résultat de la fonction objective, $F(k)$, servant de mesure d'évaluation, est ensuite calculé pour chacune d'elles, ce qui permet de d'identifier le segment du spectre le plus favorable. Une fois cette zone optimale détectée, on y teste à nouveau quelques valeurs plus rapprochées, parmi lesquelles est retenu le k dont $\arg\, max F(k)$. Cette valeur a été considérée comme suffisamment proche de l'optimum pour être retenue au sein des différents modèles LSA utilisés dans cette étude.

Concrètement, les valeurs de départ sont 10, 50, 100, 150 et 300. En effet, nous avions employé 10 dimensions durant la phase d'implémentation et de tests et avions observé d'excellents résultats. Deerwester *et al.* (1990) estiment quant à eux que 50 à 100 dimensions suffisent, tandis que Landauer (2004, cité par Bestgen, 2011, 109) recommande plutôt 300 facteurs orthogonaux. Enfin, nous avons également employé 150 dimensions afin de raffiner l'exploration de l'espace situé entre 100 et 300. Quant à la fonction objective $F(k)$, il s'agit simplement du ρ de Spearman.

Il faut noter que cette recherche a dû respecter un protocole strict afin de ne pas être sujette à un biais. En effet, comme l'objectif des variables qui découlent de la LSA est d'évaluer la cohérence de nouveaux textes, il est nécessaire de calculer la fonction objective sur un corpus de test comprenant des documents différents de ceux ayant servi à créer l'espace sémantique. Dans le cas contraire, la valeur obtenue pour ρ serait trop optimiste par rapport aux performances des variables en contexte réel. Pour éviter ce biais, les textes du *Corpus6CE* ont été scindés en un corpus de test et un d'entraînement. Les données de tests ont été obtenues par un échantillonnage stratifié qui a sélectionné aléatoirement 60 textes par niveau, cela afin de s'assurer que les probabilités *a priori* y sont bien balancées. L'ensemble des 1433 documents restants a servi pour la composition de l'espace vectoriel.

Les résultats de cette première phase de recherche, visibles sur la Figure 5.7, révèlent que la zone optimale pour k est comprise entre 0 et 50. Nous avons dès lors exploré les valeurs supplémentaires 5, 15, 20 et 40, dont les résultats sont également rapportés sur la Figure 5.7.

Il en ressort que le nombre idéal de dimensions est très petit en comparaison des valeurs rapportées dans la littérature, puisque seuls 15 facteurs orthogonaux suffisent à obtenir un ρ de $0,626$ pour **avLocalLsa_Lem**. Les résultats correspondant à ce k sont repris dans le Tableau 5.22.

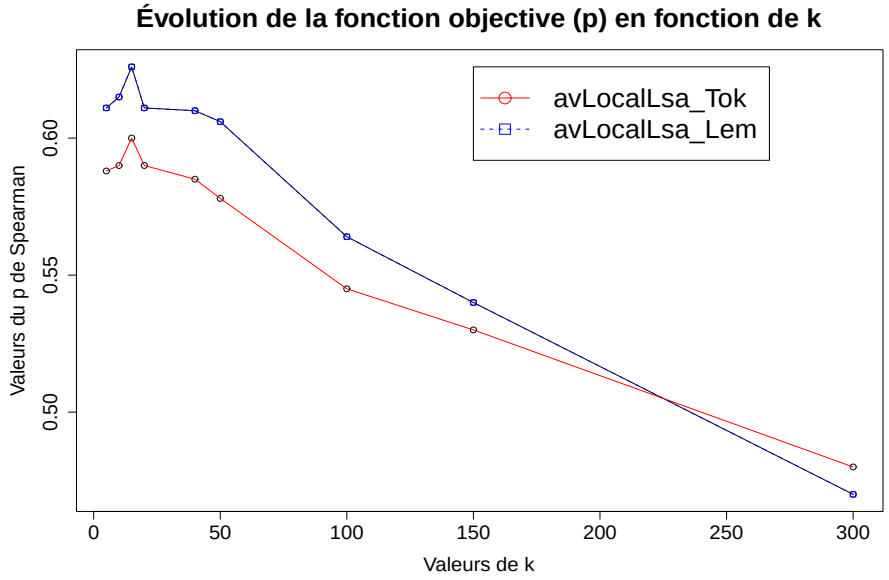


*Figure 5.7 – Valeurs de ρ en fonction de k pour **avLocalLsa_Tok** et **avLocalLsa_Lem**, calculées sur le Corpus6CE.*

Cette opération de sélection de l'espace sémantique adéquat terminée, il a été possible de comparer l'efficacité des quatre prédicteurs repris dans cette famille. Les deux premiers mesurent la cohérence directement à partir des vecteurs numériques de chaque phrase, modifiés en fonction du *tf-idf* calculé sur le corpus d'entraînement. On note qu'ils obtiennent tous deux des corrélations non significatives avec la difficulté. Soulignons que ces mesures correspondent également à l'approche LSA où k vaudrait 1433, c'est-à-dire où il n'y a pas de réduction de dimensions. Cela confirme donc le fait que la réduction de dimensions améliore considérablement les performances des prédicteurs de cette famille. En effet, le meilleur d'entre eux, **avLocalLsa_Lem**, qui prend en compte les lemmes et non les formes fléchies, atteint un ρ de $0,63$ et devient ainsi la seconde meilleure variable de notre ensemble.

Il s'agit là d'un résultat opposé à celui de Pitler et Nenkova (2008), qui rapportent un r de $-0,1$, non significatif, pour la LSA appliquée à leurs données. Toutefois, leur article ne dit rien du nombre de dimensions retenues et applique la technique sur un corpus dont le spectre de difficulté est plus réduit que le nôtre. Cependant, il pourrait bien y avoir une autre explication à cette divergence. En effet, le sens de la corrélation observée sur notre corpus est positif, ce qui signifie que plus un texte est cohérent, plus il est difficile à lire. Ce constat

s'oppose évidemment avec l'hypothèse théorique qui fonde l'usage de la LSA en lisibilité. Néanmoins, il n'est pas complètement inattendu, car les textes pour débutants comprennent davantage d'énoncés peu connectés avec leur environnement. En particulier, rappelons que notre corpus comprend 94 documents constitués de phrases isolées (*cf.* Tableau 4.3), situés à 82% dans les niveaux A1, A2 et B1. Il nous semble toutefois que le nombre de ces cas ne permet pas d'expliquer une corrélation aussi élevée et qu'il convient de conclure à une plus grande cohérence des textes utilisés dans les niveaux avancés des manuels de FLE. Ce fait s'explique probablement par la plus grande présence de documents authentiques et à destination de natifs à ces niveaux. La LSA aurait donc une pertinence toute particulière dans le contexte spécifique du FLE, ce qui constituerait une découverte d'intérêt. La présence des textes « phrases » rend néanmoins nécessaire la tenue d'expériences supplémentaires avant de conclure de manière définitive sur ce point.

	Test6CE				Test9CE			
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
coRefLocal1_Tok	$-0,04$	$0,08$	$< 0,001$	$0,63$	$-0,10^{1}$	$-0,005$	$< 0,001$	$0,35$
coRefLocal1_Lem	$-0,1$	$0,04$	$< 0,001$	$0,54$	$-0,07$	$0,001$	$< 0,001$	$0,42$
avLocalLsa_Tok	0.61^{3}	$0,6^{3}$	$< 0,001$	$0,05$	0.53^{3}	$0,54^{3}$	$< 0,001$	$0,36$
avLocalLsa_Lem	$\mathbf{0,63}^{3}$	$\mathbf{0,63}^{3}$	$< 0,001$	$0,01$	$\mathbf{0,57}^{3}$	$\mathbf{0,57}^{3}$	$< 0,001$	$0,05$

Tableau 5.22 – Taux de cohérence et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : $^{1} : < 0.05$; $^{2} : < 0.01$; $^{3} : < 0.001$.

Conclusion

Au terme de cette discussion des variables sémantiques, il convient de souligner que l'efficacité de plusieurs d'entre elles a été confirmée sur nos données, alors que les considérations rapportées au Chapitre 3.2.5 avaient donné lieu à des sérieux doutes sur ce point. Nos expérimentations ont pourtant permis d'identifier au moins trois facteurs entretenant une corrélation moyenne ou élevée avec la difficulté des textes, à savoir **PP2**, **PP1P2** et **avLocalLsa_Lem**. Si l'efficacité de ces trois facteurs était déjà attestée pour l'anglais et, en ce qui concerne la proportion de pronoms personnels, également pour le français, aucune expérience n'avait été tentée pour le FLE.

En particulier, il s'avère que la technique de la LSA détecte efficacement les différences de niveau entre des textes issus de manuels, bien que ne soit pas pour les raisons avancées dans les travaux sur la L1. Il y a donc bien un effet particulier du contexte d'application sur l'efficacité de cette dernière variable. Plus globalement, ces résultats confirment la possibilité d'effectuer, dans le contexte du FLE aussi, une synthèse entre les approches structuro-

cognitive et classique, ce qui constitue l'un des objectifs de cette thèse.

5.2.7 Analyse des variables spécifiques au FLE

La dernière famille d'indices traite d'aspects spécifiques au contexte de la lisibilité des textes de FLE, notamment en tenant compte des différences entre les processus de lecture en L1 et en L2. Comme annoncé précédemment (*cf.* Chapitre 3.2.6), seuls deux de ces aspects ont pu être pris en compte dans cette étude, essentiellement à cause d'un manque de données adéquates à une approche statistique. Il s'agit tout d'abord des UPs, supposés représenter un problème spécifique pour les apprenants et dont diverses caractéristiques ont été considérées, et, dans un second temps, d'un ensemble de variables destinées à détecter la présence de dialogues au sein des textes à évaluer.

Les unités polylexicales

Dans le Chapitre 3.2.6, il a été avancé que les UPs sont susceptibles d'engendrer des difficultés à la lecture chez des apprenants de FLE. Ces difficultés pourraient simplement s'expliquer en fonction de leur densité dans les textes, mais aussi, plus subtilement, en fonction de variations de leur répartition au sein des différents niveaux de difficulté, variations qui sont mesurées soit via leur fréquence, soit via la proportion relative de leurs différentes structures. Plusieurs expérimentations ont été menées afin d'éclaircir ces questions, qui n'ont pas encore été vraiment traitées dans la littérature [8].

Comme expliqué au Chapitre 3.3.4, la paramétrisation des variables basées sur les UPs nominales fait usage de deux bases lexicales – *Le Soir* et Google –, requises pour estimer le score d'association attribué à chaque candidat UPs. De plus, le seuil à partir duquel un candidat a été considéré comme une UPs, noté θ, a également fait l'objet d'expérimentations. Quatre valeurs ont été retenues pour chaque référence : un seuil nul, où toutes les structures lexicales nominales sont considérées ; un second et un quatrième seuil qui correspondent respectivement à une précision de 30% et 50% pour notre extracteur [9] et, comme troisième seuil, une valeur intermédiaire entre ces deux niveaux de précision. De cette manière, il est possible d'analyser le comportement des 44 variables basées sur les UPs à mesure que la force d'association liant les constituants des UPs augmente. Par contre, la multiplication des expériences a, une fois de plus, conduit à une explosion des résultats. Par conséquent, seules les corrélations de Spearman pour le *Corpus6CE* sont présentées dans le Tableau

8. Comme pour les n-grammes, les résultats rapportés ici ont fait l'objet d'études préalables, réalisées en collaboration avec Patrick Watrin et publiées par ailleurs (François et Watrin, 2011a,b).

9. Pour les détails de la procédure d'évaluation employée pour calculer ces taux de précision, se reporter à Watrin et François (2011).

C.2, rendant ainsi plus facile à suivre leur évolution en fonction des quatre seuils et des deux références.

	Le Soir				Google			
Seuils	0	19	25	43	0	139	4000	9931
NCPW	$\mathbf{0.242}^{3}$	$\mathbf{0.183}^{3}$	$\mathbf{0.146}^{2}$	$\mathbf{0.135}^{2}$	$\mathbf{0.199}^{3}$	$\mathbf{0.16}^{3}$	$\mathbf{0.178}^{3}$	$\mathbf{0.199}^{3}$
MeanSize	0.055	0.086	0.066	0.061	-0.017	-0.016	0.008	-0.013
NNColl	0.053	0.167^{3}	0.18^{3}	0.157^{3}	0.052	0.073	0.053	0.045
NPNColl	0.053	0.115^{1}	0.125^{2}	$\mathbf{0.173}^{3}$	0.037	0.163^{3}	0.21^{3}	$\mathbf{0.214}^{3}$
ANColl	0.006	0.069	0.079	0.071	-0.003	0.109^{1}	0.141^{2}	0.142^{3}
NAColl	$\mathbf{0.283}^{3}$	$\mathbf{0.282}^{3}$	$\mathbf{0.271}^{3}$	$\mathbf{0.286}^{3}$	$\mathbf{0.31}^{3}$	$\mathbf{0.325}^{3}$	$\mathbf{0.262}^{3}$	$\mathbf{0.253}^{3}$
MeanPColl	0.116^{2}	0.077	-0.066	0.05	0.254^{3}	0.18^{3}	0.109^{1}	0.098
MedianPColl	-0.083	-0.108^{1}	-0.075	-0.044	0.032	0.047	0.047	0.062
PColl75	-0.225^{3}	-0.261^{3}	-0.202^{3}	-0.15^{1}	-0.131^{2}	-0.064	-0.022	0.008
PColl90	$\mathbf{-0.375}^{3}$	$\mathbf{-0.367}^{3}$	$\mathbf{-0.296}^{3}$	$\mathbf{-0.207}^{3}$	$\mathbf{-0.341}^{3}$	$\mathbf{-0.17}^{2}$	-0.05	-0.006
MeanP2Coll	0.078	0.028	0.001	-0.028	0.203^{3}	0.09	0.054	0.037
MedianP2Coll	-0.103^{1}	-0.069	-0.096	-0.073	-0.088^{1}	-0.034	0.018	0.044
P2Coll75	-0.236^{3}	-0.172^{3}	-0.188^{3}	-0.15^{1}	-0.245^{3}	-0.114^{1}	-0.029	0.022
P2Coll90	$\mathbf{-0.284}^{3}$	$\mathbf{-0.241}^{3}$	$\mathbf{-0.251}^{3}$	$\mathbf{-0.21}^{3}$	$\mathbf{-0.283}^{3}$	$\mathbf{-0.162}^{2}$	-0.055	0.005
MeanP3Coll	0.093	0.0	-0.046	-0.018	0.168^{3}	0.1	0.1	0.057
MedianP3Coll	-0.167^{3}	-0.086	-0.101	-0.064	-0.083	0.075	0.094	0.059
P3Coll75	-0.235^{3}	-0.163^{2}	-0.158^{2}	-0.106	-0.174^{3}	0.045	0.081	0.045
P3Coll90	$\mathbf{-0.242}^{3}$	$\mathbf{-0.183}^{3}$	$\mathbf{-0.192}^{3}$	-0.131^{1}	$\mathbf{-0.183}^{3}$	0.025	0.069	0.039
MeanP4Coll	0.12^{1}	0.044	-0.051	-0.146	0.167^{2}	0.084	0.035	0.09
MedianP4Coll	-0.098	-0.025	-0.058	-0.145	-0.101^{1}	0.074	0.029	0.083
P4Coll75	-0.123^{1}	-0.033	-0.097	-0.167	-0.171^{3}	0.054	0.021	0.077
P4Coll90	-0.124^{1}	-0.039	-0.128	-0.181	-0.184^{3}	0.039	0.014	0.071

Tableau 5.23 – Variables d'UPs et difficulté : ρ de Spearman pour les deux corpus et les 4 seuils. Le taux de significativité des corrélations est indiqué comme suit : $^{1} : < 0.05$; $^{2} : < 0.01$; $^{3} : < 0.001$.

À la lecture de ce tableau, les variables qui ressortent sont : (1) la densité des UPs (**NCPW**) ; (2) les 75e et 90e percentiles des distributions de fréquences (**PColl90**, **P2Coll90**, etc.), qui, comme c'était déjà le cas pour les variables lexicales basées sur la fréquence, se comportent nettement mieux que la moyenne et la médiane et (3) la proportion de deux structures en particulier : **NAColl** et, dans une moindre mesure, **NPNColl**.

On notera que, de manière générale, la plupart des variables perdent en capacité de discrimination à mesure que θ augmente. Cette situation est particulièrement problématique. En effet, si la totalité des constituants nominaux est considérée lorsque $\theta = 0$, le renforcement de la cohésion des unités, qui va de pair avec un accroissement de la précision de l'extracteur, signifie qu'une proportion de plus en plus grande de véritables UPs sont obtenues via le processus d'extraction. Or, il apparaît que cette hausse de la précision ne conduit pas à une amélioration des performances des variables basées sur les UPs, contrairement à l'hypothèse posée pour cette famille de prédicteurs.

Ce constat, similaire à celui obtenu précédemment (François et Watrin, 2011a), peut cepen-

dant s'expliquer par au moins deux raisons (*cf.* aussi le Chapitre 3.2.6). Soit l'hypothèse théorique selon laquelle la fréquence des UPs influence la lecture est inadéquate. Malheureusement, notre expérience ne permet pas de prendre position vis-à-vis de cette question, étant donné que c'est la répartition des UPs dans les textes qui est mesurée et non le taux de compréhension des UPs. Une autre possibilité est que l'extracteur d'UPs ne parvient pas à identifier correctement les UPs qui sont sources de difficulté. En effet, parmi les meilleurs candidats pour notre corpus, on trouve des UPs telles que « effet de serre » ou « développement durable », pertinentes dans un contexte d'apprentissage des L2, mais aussi « mardi soir » ou « millions d'euros », qui ne comportent aucune difficulté de compréhension. Dès lors, il serait nécessaire, pour valider plus strictement cette hypothèse, de repérer manuellement les UPs, en ne considérant que celles qui sont sémantiquement opaques. Cette perspective sort cependant du cadre de notre étude.

Une troisième source de problèmes pour ces variables pourrait se situer dans la nature des références. On observe ainsi des différences notables entre les résultats obtenus à partir des n-grammes de Google et du *Soir*. D'une part, les meilleurs facteurs présentent des corrélations plus élevées à partir du *Soir*, à l'exception de **NAColl**. D'autre part, il apparaît que tous les indices basés sur des chaînes nominales de plus de deux mots sont non significativement corrélés avec la difficulté dans le cas de Google. Or, cette base comporte plusieurs défauts que n'a pas *Le Soir*, dont une tokenization poussée à l'extrême et un vocabulaire étrangement réduit en comparaison du nombre de documents indexés.

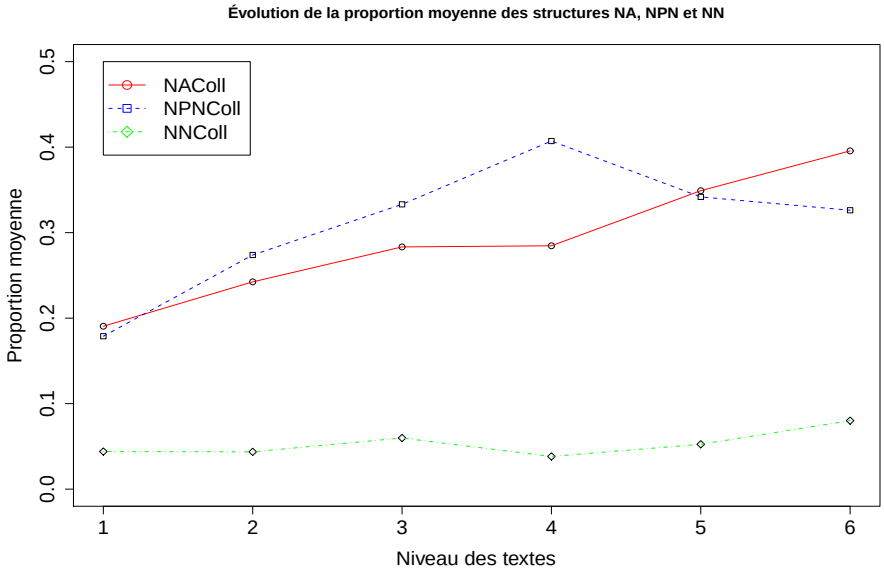


Figure 5.8 – Évolution de la proportion moyenne des structures NA, NPN et NN par niveau de difficulté. Obtenu à partir d'un des dix échantillons de Corpus6CE ayant servi à calculer les corrélations.

Insistons toutefois sur trois exceptions à cette baisse d'efficacité en relation avec θ, qui, toutes trois, sont liées aux types de structures que présentent les UPs. Il s'agit de **NAColl**, dont la proportion dans les textes continue à apporter une information intéressante quelle que soit la valeur de θ, ainsi que de **NNColl** et **NPNColl**, dont la capacité de discrimination augmente de pair avec θ. Ce constat reproduit et approfondit les résultats de notre étude précédente (François et Watrin, 2011b), dont l'une des conclusions était que le type de structure des UPs possèdent bien un pouvoir discriminant, à la différence de **MeanPColl** dont l'inefficacité réplique les résultats de Ozasa *et al.* (2007) sur l'anglais. Le graphique 5.8 montre, pour l'un des échantillons de *Corpus6CE*, les proportions moyennes par niveau de ces trois structures pour $\theta = 43$ (la référence est ici *Le Soir*). On note que si **NAColl** présente un profil monotone croissant, ce n'est pas le cas de NPNColl, dont les moyennes sont plutôt bien différenciées, mais pas situées sur une courbe ascendante. Il n'est dès lors pas étonnant que ce facteur soit l'un des seuls de cette famille pour lequel l'hypothèse de linéarité ($p - value = 0,037$) est rejetée.

En conclusion, nos expérimentations sur l'intérêt de prendre en compte les UPs au sein d'une formule de lisibilité démontrent que celui-ci est très faible. D'autant que des expériences de modélisation effectuées précédemment (François et Watrin, 2011a) ont montré qu'elles n'apportent pas d'informations supplémentaires, ni par rapport à une « baseline » classique, ni par rapport à une simple variable de fréquence prenant en compte tous les n-grammes du texte au lieu de se restreindre aux UPs détectées automatiquement. Il a cependant bien fallu opter pour un seuil donné pour les étapes suivantes de la modélisation. Bien que les seuils les plus élevés (43 et 9 931) soient les moins efficaces, ils ont été retenus dans ce but afin de rester en accord avec la définition de cette famille de variable.

Type de texte

Terminons cette analyse des différents prédicteurs par les variables, proposées par Henry (1975), destinées à détecter le type de texte ou, plus exactement, la présence de dialogues au sein des textes à évaluer. Cette information nous avait paru susceptible de discriminer efficacement le niveau des textes en FLE étant donné que les documents employés dans les niveaux débutants prennent souvent la forme de conversations. Cette hypothèse se voit confirmée, dans une certaine mesure, par les corrélations du Tableau 5.24. Deux facteurs ressortent en effet : **PPD** et **BINGUI**.

Pour **PPD**, ce résultat n'est pas surprenant, vu qu'il s'agit en réalité d'une variable très similaire à **PP1P2**, qui s'est avéré le meilleur des indices du taux de personnalisation des textes (*cf.* Section 5.2.6). La différence entre les deux provient uniquement de l'ajout des pronoms possessifs de la 1re et de la 2e personne. D'ailleurs, ces deux facteurs sont intercorrélés à

	Test6CE				Test9CE			
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
PPD	$\mathbf{-0,396}^3$	$\mathbf{-0,318}^3$	$< 0,001$	$0,026$	$\mathbf{-0,401}^3$	$\mathbf{-0,337}^3$	$< 0,001$	$< 0,001$
PI	$-0,284^3$	$-0,245^3$	$< 0,001$	$< 0,001$	$-0,309^3$	$-0,269^3$	$< 0,001$	$< 0,001$
PPEI_1	$-0,212^3$	$-0,103$	$< 0,001$	$< 0,001$	$-0,253^3$	$-0,166^3$	$< 0,001$	$0,002$
PPEI_2	$-0,261^3$	$-0,139^1$	$< 0,001$	$0,001$	$-0,291^3$	$-0,199^3$	$< 0,001$	$< 0,001$
BINGUI	$\mathbf{0,462}^3$	$\mathbf{0,462}^3$	$< 0,001$	$0,018$	$\mathbf{0,45}^3$	$\mathbf{0,45}^3$	$< 0,001$	$0,311$

Tableau 5.24 – Indices du type de texte et difficulté : r de Pearson, ρ de Spearman, p-valeur du test de Shapiro-Wilk ($W(p)$) et p-valeur du test F de Guilford ($F(p)$). Le taux de significativité des corrélations est indiqué comme suit : $^1 : < 0.05$; $^2 : < 0.01$; $^3 : < 0.001$.

hauteur de $\rho = 0,98$, ce qui rend inutile leur introduction conjointe au sein d'un même modèle.

Quant à **BINGUI**, il est étonnant de constater que la seule présence de guillemets de dialogue dans le texte permet d'atteindre une corrélation de $0,462$ avec le critère, alors que la ponctuation exclamative et interrogative, ainsi que la présence d'interjections apportent une information moins utile. Par ailleurs, **BINGUI** semble bien apporter une information spécifique, comme le montre les différentes corrélations croisées rapportées au Tableau 5.25. Il s'agit donc *a priori* d'un nouveau prédicteur intéressant.

	ML3	NMP	PPD	PI	PPEI_1	avLocalLsa_Lem
BINGUI	$-0,18$	$0,35$	$-0,17$	$-0,01$	-0.05	$0,4$

*Tableau 5.25 – Corrélations croisées (ρ de Spearman) entre **BINGUI** et quelques variables typiques, calculées sur l'un des 10 échantillons issus de Corpus6CE, sélectionné au hasard.*

Néanmoins, le fait que la corrélation soit positive entre la difficulté des textes et **BINGUI** interpelle. Il s'agit à nouveau d'un résultat contradictoire avec l'hypothèse de départ. En analysant les données de plus près, ce phénomène semble s'expliquer en partie par un effet de longueur des textes (ρ entre les deux variables est de $0,48$ sur un échantillon de *Corpus6CE*) : plus ceux-ci sont longs et plus il y a de chance qu'ils contiennent des guillemets de dialogue. Cette dépendante possible par rapport à la longueur des textes explique aussi la corrélation relativement bonne obtenue par **BINGUI**, mais diminue par contre la validité de ce prédicteur. Une solution pour supprimer cette dépendance constituerait à calculer la proportion de mots entre des guillemets de dialogue, normalisée par la longueur du texte. Nous avons toutefois laissé cette expérience comme perspective pour de futurs travaux sur la question.

Conclusion : les variables spécifiques au FLE

Si l'analyse des variables lexicales, syntaxiques et sémantiques a mis au jour divers candidats adaptés au contexte du FLE, les facteurs spécifiques à ce contexte que nous avions identifiés n'ont guère tenu leurs promesses. En ce qui concerne les mesures basées sur les UPs, il peut s'agir d'un problème de reconnaissance de ces entités comme d'une faiblesse au niveau de l'hypothèse concrétisée par ces variables. Quant aux variables de dialogue, elles semblent avoir un certain pouvoir discriminatif, mais la meilleure d'entre elles, **BINGUI**, semble en réalité dépendante de la longueur des textes.

Dès lors, cette famille est certainement celle qui a produit les moins bons prédicteurs. Il faut cependant rappeler que la plus pertinente des dimensions spécifiques au processus de lecture en L2, à savoir l'effet facilitateur des congénères, n'a pas pu être investiguée par manque de données adaptées. De plus, l'intérêt scientifique des expériences réalisées dans cette section est loin d'être nul, étant donné que nous n'avons pas connaissance d'une analyse similaire aussi poussée dans la littérature, tant pour les variables d'UPs que pour celles de dialogue. Il y a donc matière à diverses perspectives, telles qu'une mesure du ratio de texte dialogué ou des variables prenant en compte uniquement les UPs sémantiquement opaques. Cette dernière piste requerrait cependant le développement d'un extracteur capable de distinguer les expressions dont le sens est non-compositionnel. Or, les réalisations actuelles dans le domaine, utilisant soit la substitution de mots similaires (Lin, 1999), soit des techniques statistiques (Baldwin *et al.*, 2003), restent encore trop imprécises pour une telle tâche.

5.2.8 Conclusion : quelle efficacité pour nos prédicteurs ?

Au terme de cette longue analyse des 406 prédicteurs à notre disposition, plusieurs faits intéressants peuvent être relevés. Le plus important d'entre eux est sans doute le fait qu'il ait été possible d'identifier, pour chaque dimension linguistique, au moins un prédicteur efficace. Il s'agit de **PA_Alterego1a**, au niveau lexical ($\rho = 0,652$) ; de **NMP**, pour les aspects syntaxiques ($\rho = 0,618$) ; de **avLocalLsa_Lem** pour la sémantique ($\rho = 0,63$) et, dans une moindre mesure, de **BINGUI** en ce qui concerne les aspects spécifiques au FLE ($\rho = 0,46$). Cela confirme donc notre hypothèse selon laquelle il devait être possible d'identifier de nouveaux prédicteurs pour le FLE au sein de nouvelles dimensions textuelles et laisse présager de la possibilité de dépasser les performances atteintes par une formule de type classique sur nos données grâce à ces nouvelles variables.

Toutefois, on note aussi que, pour les facteurs lexico-syntaxiques, les prédicteurs obtenus sont tout à fait classiques : le nombre d'absents d'une liste de mots faciles et le nombre moyen de mots par phrase. Si ce n'était l'efficacité de la LSA, il serait possible de conclure

que c'est avec les vieilles casseroles qu'on fait les meilleures formules de lisibilité, ce qu'avait d'ailleurs déjà observé Bormuth (1969a) à propos des indices syntaxiques.

Le fait que la LSA concurrence ces prédicteurs classiques ouvre toutefois des perspectives intéressantes. Cela montre d'une part que l'emploi de techniques de TAL ne conduit pas nécessairement à une impasse, comme aurait pu le laisser croire la faible efficacité des variables basées sur les modèles n-grammes ou sur les UPs. Ce dernier exemple révèle en particulier que l'automatisation de notions linguistiques plus fines ne se traduit pas nécessairement par un gain d'information, probablement parce que les traitements de TAL requis comportent trop d'approximations (dans ce cas, des erreurs de l'extracteur d'UPs, des problèmes de couverture des références, etc.). Par contre, le succès de la LSA confirme qu'il est effectivement possible de concevoir des prédicteurs efficaces recourant au TAL. Cela ouvre donc la voie à une approche computationnelle de la lisibilité du FLE, qui, à l'instar des travaux de Pitler et Nenkova (2008), vise à proposer une synthèse des approches classique et structuro-cognitive.

Un dernier aspect remarquable concerne la distribution des prédicteurs. Nous l'avions souligné dans la conclusion sur les prédicteurs lexicaux (*cf.* Section 5.2.4), mais l'étude des autres aspects textuels a confirmé que la grande majorité de nos variables ne peut être considérée comme gaussiennes, d'après le test de Shapiro-Wilk. Il s'agit là aussi d'un résultat divergeant par rapport aux travaux de Bormuth (1966), puisque même les variables issues de la moyenne de variables aléatoires non normales ne sont pas distribuées normalement. Par contre, seule une proportion d'entre elles ne sont pas associées linéairement avec la difficulté. Cette proportion est toutefois suffisante pour mettre en doute la capacité de modèles statistiques linéaires à utiliser au mieux l'information contenue dans notre jeu de données. L'emploi de la régression linéaire multiple, classiquement utilisé en lisibilité, semble donc à éviter. C'est pourquoi nous avons envisagé d'autres modèles statistiques dans la seconde partie de ce chapitre, qui décrit l'étape de conception de notre formule.

5.3 Vers un modèle de lisibilité

5.3.1 Introduction

Les expériences réalisées dans la première partie de ce chapitre ont permis d'identifier un ensemble de variables significativement associées à la difficulté des textes de notre corpus. Or, comme le carré du r de Pearson représente aussi la proportion de variance qu'un facteur X explique au sein de la distribution d'un phénomène Y, on peut considérer qu'une variable présentant une corrélation élevée avec le niveau des textes se révèlera un indice de lisibilité efficace. Ce tableau optimiste doit toutefois être nuancé étant donné que les variables simi-

laires, en particulier celles appartenant à une même famille, sont susceptibles d'apporter une information identique. Dans ce cas, l'une des deux devient alors parfaitement inutile.

C'est la découverte de ce phénomène qui a motivé les auteurs de la période classique à limiter leurs formules à deux prédicteurs, une position ensuite remise en question par Bormuth (1966, 1969a), qui défend la conception de modèles plus complexes. C'est pourquoi nous débuterons cette seconde partie, qui traite des aspects méthodologiques propres à la conception de la meilleure formule de lisibilité possible pour le FLE, par une discussion du nombre de variables que celle-ci devrait idéalement contenir (*cf.* Section 5.3.2), avant de détailler les deux techniques mises en place pour opérer une sélection parmi les 406 variables décrites dans le Chapitre 3.

Une fois les variables sélectionnées, il convient de déterminer lesquelles sont les plus pertinentes en fonction de l'objectif à atteindre. Cela revient à attribuer un poids à chaque facteur retenu dans la formule, afin de rendre celle-ci capable de prédire plus ou moins correctement la difficulté de nouveaux textes. Cette étape de pondération est envisagée dans un second point (*cf.* Section 5.3.3), où seront brièvement expliqués les différents algorithmes statistiques employés à cet effet. Une fois ces outils définis, il sera possible de comparer les différents modèles qui en découlent, ce qui est fait dans la troisième et dernière partie de ce chapitre.

5.3.2 Nombre de variables dans une formule et stratégie de sélection

À l'origine, la question du nombre de variables à retenir dans une formule est surtout dirigée par des considérations pratiques. Comme la paramétrisation s'effectue manuellement, mieux vaut limiter la quantité de facteurs. Des auteurs tels que Gray et Leary (1935) n'hésitent cependant pas à en employer jusqu'à cinq au sein de leur modèle. Durant l'époque classique, plusieurs travaux (Flesch, 1948; Lorge, 1944a) défendent l'opinion selon laquelle il ne servirait à rien de développer des formules comportant plus de deux ou trois variables, car la majorité de celles-ci sont hautement intercorrélées.

Ce point de vue semble toujours dominer trente ans plus tard, lorsque Klare (1974, 96) avance, dans sa revue du domaine :

> *Unless a user is interested in doing research, there is little to be gained from choosing a highly complex formula. A simple 2-variable formula should be sufficient, especially if one of the variables is a word or semantic variable and the other is a sentence or syntactic variable. Beyond these 2 variables, further additions add relatively little predictive validity compared to the added application time involved, and a formula with very many variables is likely to be unreliably*

> *applied by hand.*

De cette citation, on peut retirer deux enseignements supplémentaires. D'une part, que la conception classique se justifierait dans la majorité des cas parce que l'apport informationnel devient trop faible à partir de la troisième variable, ce qui a effectivement été observé dans la plupart des travaux de l'époque (*cf.* Chapitre 2). Cela explique pourquoi il s'avère difficile de surpasser les modèles classiques, que l'on pourrait pourtant qualifier de « baseline ». D'autre part, Klare estime que les variables insérées dans un modèle devraient traiter de dimensions textuelles différentes, ce qui est aussi notre conviction.

Toutefois, certains auteurs ne s'accordent pas avec cette conception classique. Ainsi, Bormuth (1966, 112-113) effectue des expériences où les variables du modèle sont introduites l'une après l'autre et il observe une progression des performances bien au-delà du second facteur. Or, les prédicteurs qu'il ajoute ainsi avec succès sont innovants et concernent des aspects peu ou prou couverts précédemment. Bormuth (1966, 113) en conclut :

> *It is apparent that available readability formulas contain too few variables and thereby ignore variables of prose style which can make important independent contributions to the multiple correlations.*

Il estime également qu'un nombre trop limité de facteurs augmente le risque de mal traiter les textes présentant des caractéristiques particulières, dont un exemple extrême est un texte où tous les mots sont mélangés.

Par la suite, Bormuth (1969a) décèle tout de même un problème de surapprentissage pour les formules comportant un grand nombre de variables (de 14 à 20). Leur coefficient de corrélation multiple, R, diminue en effet significativement en validation croisée. Il y aurait donc un nombre optimal de facteurs à intégrer : ni trop, ni trop peu. Dès lors, sous l'effet de l'informatisation croissante, le souci de facilité d'application des formules laisse peu à peu place à une autre préoccupation : la notion de parcimonie [10]. Il importe alors davantage de définir un modèle comportant le moins de variables possibles, afin de maximiser les performances en validation croisée, une étape souvent omise dans les études de l'époque classique. C'est cette approche, initiée par Bormuth (1966) au sein de la lisibilité, que nous suivrons dans cette étude.

Pour notre étude, deux approches différentes ont été utilisées pour essayer de sélectionner le meilleur ensemble de facteurs : l'approche experte et l'approche automatisée. L'approche

10. Celle-ci est attribuée au philosophe nominaliste du 14e siècle, Guillaume d'Ockham, qui a insisté sur le fait que, parmi un ensemble d'explications possibles pour un phénomène, la plus correcte est la plus simple (Crawley, 2007, 325). Pour cette raison, on désigne parfois aussi la notion de parcimonie à l'aide de l'epxression le « rasoir d'Ockham ». En termes de modélisation statistique, cela revient à limiter au minimum le nombre de variables explicatives, à privilégier des modèles linéaires à des modèles non linéaires, etc.

experte consiste à se baser sur les analyses effectuées dans la première partie de ce chapitre pour sélectionner un ensemble de variables efficaces. Elle présente l'avantage de limiter sévèrement le nombre de prédicteurs et de reposer sur une expertise. Par contre, comme nous n'avons pas envisagé systématiquement toutes les corrélations croisées entre variables, le risque d'en sélectionner certaines apportant une information similaire augmente et cette stratégie n'est, en théorie, pas optimale. Nous avons néanmoins défini trois ensembles de prédicteurs selon cette méthode :

Baseline : NMP + NLM
: Elle correspond aux formules classiques et représente le résultat à améliorer.

Expert1 : PA_Alterego1a + NMP + av_LocalLsa_Lem + BINGUI
: Considère la meilleure variable au sein de chacune des quatre grandes familles

Expert2 : PA_Alterego1a + X90FFFC + NMP + PPres + av_LocalLsa_Lem + PP1P2 + BINGUI + NAColl
: Considère cette fois les deux meilleures variables par dimension linguistique

La seconde manière de constituer un ensemble de variables consiste à les sélectionner automatiquement à l'aide de techniques reposant sur des critères statistiques. Déjà utilisée par Bormuth (1966), ce type de stratégie présente l'avantage de prendre en compte les intercorrélations entre variables et de maximiser la quantité d'information présente au sein du modèle. Par contre, elle est généralement fort sensible aux données d'entraînement : nous avons ainsi observé lors d'expériences préparatoires qu'une même procédure de sélection parvenait à des conclusions différentes selon qu'elle s'appliquait au *Corpus6CE* ou au *Corpus9CE* [11]. C'est pourquoi nous avons tenu à comparer ces méthodes, qui font partie de l'état de l'art actuel en intelligence artificielle, à une approche experte. Leurs détails d'implémentation variant d'un algorithme statistique à l'autre, ils sont discutés dans les sections se rapportant à chacun de ces algorithmes.

5.3.3 La pondération des variables

Introduction

Une seconde problématique dans la conception d'une formule de lisibilité consiste à préciser l'importance relative de chaque facteur retenu au sein du modèle. La solution la plus simple consisterait à attribuer des poids à chacun d'eux sur la base de postulats théoriques. Malheureusement, les chercheurs en lisibilité ne sont jamais parvenus à s'accorder sur une telle hiérarchie. D'une étude à l'autre, c'est tantôt le lexique, tantôt une mesure de la com-

11. Cette sensibilité des procédures de sélection automatique des variables aux données est déjà remarquée par Bormuth (1966, 108) qui indique : « With slight chance variations in the data, another set of variables could have entered the equation ».

plexité syntaxique, tantôt encore une variable cognitive qui se révèle la plus corrélée avec le critère.

Comme premier exemple de ce désaccord, citons le débat qui a eu lieu, dans la seconde moitié des années 40, entre Flesch (1946) et Dale et Chall (1948a). Depuis les années 20, les mesures lexicales étaient considérées comme primordiales, mais, en 1943, Flesch remet en question cette domination. Il estime que la reconnaissance des mots constitue essentiellement un problème pour les lecteurs débutants ou faibles, mais que pour une formule généraliste comme la sienne, c'est surtout le niveau d'abstraction et la syntaxe qui importent. Toutefois, Dale et Chall (1948a, 17) observent au contraire, sur la base des corrélations qu'ils ont calculées, que le lexique reste bien le facteur prédominant, comme l'affirmait déjà Lorge (1944a).

Ce débat se poursuit par la suite. Ainsi, Spache (1953, 411) confirme cette supériorité du nombre moyen de mots par phrase sur le pourcentage de mots difficiles, résultat qui est ensuite répliqué par Hull (1979, cité par DuBay, 2004, 51). Toutefois, les travaux récents de Stenner *et al.* (1983), Schwarm et Ostendorf (2005) et Heilman *et al.* (2007) concluent quant à eux à la supériorité des indices lexicaux. Enfin, Björnsson (1983, 483) remarque que les coefficients de la régression linéaire multiple varient en fonction du corpus d'entraînement utilisé et décide en conséquence d'accorder la même importance à ces deux facteurs.

En résumé, il est clair qu'une approche reposant sur des connaissances universellement valables ne peut être appliquée pour pondérer les facteurs en lisibilité. Leur importance dépend en effet des données d'apprentissage, mais aussi de la population ciblée par la formule. On trouve d'ailleurs de nombreuses attestations de ce second fait dans la littérature. Déjà Gray et Leary (1935, 16) obtenaient des coefficients différents pour certains de leurs facteurs selon qu'ils considéraient uniquement les lecteurs faibles ou avancés. Par la suite, Harris et Jacobson (1974, 3) signalent que si le pourcentage d'absents obtient la meilleure corrélation pour les *grades* 1 à 6, c'est le nombre de mots par phrase qui se comporte le mieux lorsqu'on ne considère que les *grades* 1 à 3.

On pourrait multiplier les exemples en ce sens (Dale et Tyler, 1934; Fry, 1977; Henry, 1975; Kemper *et al.*, 1993; Kintsch et Vipond, 1979; Meyer, 1982; Miller et Kintsch, 1980), mais nous nous en dispenserons, car cette problématique n'est pas directement pertinente pour notre cas. En effet, rappelons que l'objectif de notre formule est de classer des textes de FLE en fonction d'un ensemble de documents de référence et que, par conséquent, la population de lecteurs n'est pas directement considérée.

Confronté à ces divergences d'opinions assez nettes en ce qui concerne l'importance des différents indices de lisibilité, nous avons opté pour une méthodologie classique dans la

conception de modèles statistiques : laisser des algorithmes d'apprentissage calculer les coefficients optimaux pour chaque variable. Il s'agit d'ailleurs de la démarche adoptée par la majorité des chercheurs de la discipline. Cette opération est généralement réalisée à l'aide du maximum de vraisemblance (Hastie *et al.*, 2001, 31-32) et revient à maximiser $P(M \mid D)$, c'est-à-dire la probabilité du modèle (M) étant donnés les documents d'entraînement (D).

Il existe toutefois un grand nombre d'algorithmes différents, qui divergent par les techniques qu'ils emploient et par les hypothèses qu'ils imposent sur les données. En particulier, réapparaît ici la question de l'échelle de mesure employée pour la difficulté. En effet, comme il l'a été expliqué précédemment (*cf.* Section 5.2.2), il est possible de mettre en rapport chaque type de données avec des tests ou des modèles statistiques appropriés. Le Tableau 5.26 détaille un ensemble de modèles intéressants qui ont été retenus pour chacune des trois échelles de mesure adaptées à la modélisation de la difficulté.

Type d'échelle	**Modèle statistique**
	Régression logistique multinomiale (régression logistique multinomiale (RLM))
	Arbres de décision
Nominale	Bagging
	Boosting
	Séparateurs à vaste marge
Ordinale	Régression logistique ordinale (régression logistique ordinale (RLO))
Intervalle	Régression linéaire multiple

Tableau 5.26 – Classification des modèles statistiques utilisés dans cette étude d'après l'échelle de mesure des variables auxquelles ils s'appliquent.

À l'exception de la régression linéaire multiple, laquelle a été écartée suite à de mauvaises performances lors d'expériences préliminaires, le principe de chacun de ces algorithmes est esquissé dans les sections qui suivent afin que le lecteur comprenne globalement leur fonctionnement. Il n'entre toutefois pas dans notre objectif de formaliser précisément ceux-ci, étant donné qu'il s'agit d'outils standards en intelligence artificielle, dont les descriptions détaillées sont disponibles dans de nombreux ouvrages de référence (Agresti, 2002; Bramer, 2007; Hastie *et al.*, 2001; Tan *et al.*, 2005).

La régression logistique

La régression logistique est une technique statistique qui modélise la relation entre une variable réponse Y *qualitative* (nominale ou ordinale) et une ou plusieurs variables explicatives $X_1, \cdots, X_p$, qui peuvent être aussi bien *qualitatives* que *quantitatives*. Dans le modèle de base, la variable Y est dichotomique : elle prend la valeur 0 ou 1 selon que la caractéristique d'intérêt est présente ou non pour une observation donnée. Toutefois, la variable *difficulté*

telle que nous l'envisageons ne suit évidemment pas une échelle dichotomique, puisqu'elle comporte 6 ou 9 niveaux. Il existe heureusement des variantes aptes à traiter des variables qualitatives à K niveaux ($K > 2$). Celles-ci se classent en deux familles selon que la variable réponse est nominale ou ordinale : la régression logistique multinomiale ou polytomique et la régression logistique ordinale. Dans la suite de cette section, nous présenterons d'abord brièvement les principales caractéristiques du modèle logistique dichotomique, qui sont nécessaires à la compréhension des modèles plus complexes. Ensuite, nous étudierons comment la difficulté peut être modélisée selon le modèle multinomiale, avant de considérer une approche de la régression logistique ordinale basée sur la proportionnalité des rapports de chance.

La régression logistique dichotomique La régression logistique dichotomique ressemble à la régression linéaire en ce sens qu'elle modélise aussi l'espérance $E(Y \mid \mathbf{x})$, c'est-à-dire la valeur moyenne de Y pour chaque combinaison de valeurs que peuvent prendre les p variables explicatives $x_1, \cdots, x_i, \cdots, x_p$. Toutefois, elle s'en distingue par deux caractéristiques fondamentales. D'une part, comme il s'agit d'une variable binaire, $E(Y \mid \mathbf{x})$ ne peut être inférieure à 0 (on ne peut faire pire qu'un échec complet), ni supérieure à 1. D'autre part, la relation entre $E(Y \mid \mathbf{x})$ et les facteurs $x_1, \cdots, x_i, \cdots, x_p$ n'est pas linéaire. Comme le montre le graphique 5.9 pour le cas bivarié, la variation de la moyenne conditionnelle par unité de x n'est pas constant. Aux abords de 0, elle est minime, puis croît rapidement, avant de tendre vers 0 aux abords de 1. Par conséquent, il est nécessaire d'appliquer aux variables explicatives une transformation apte à reproduire cette courbe en S, que l'on appelle une sigmoïde.

En général, la fonction appliquée dans ce but est le logit de $P(Y = 1 \mid X = \mathbf{x})$ [12]. Or, cette quantité revient à :

$$logit[P(Y = 1 \mid X = \mathbf{x})] = \ln \frac{P(Y = 1 \mid X = \mathbf{x})}{1 - P(Y = 1 \mid X = \mathbf{x})} \tag{5.7}$$

Cela signifie que c'est la relation entre ce ratio et la combinaison linéaire des prédicteurs $x_1, \cdots, x_i, \cdots, x_p$ qui est modélisée, ce qui donne l'équation de régression suivante (Agresti, 2002, 183) :

$$logit[P(Y = 1 \mid X = \mathbf{x}] = \alpha + \beta_1 x_1 + \cdots + \beta_i x_i + \cdots + \beta_p x_p \tag{5.8}$$

Les valeurs des coefficients β_i sont obtenues par maximum de vraisemblance (Agresti, 2002, 192-195) et elles s'interprètent comme suit : $\exp(\beta_i)$ représente l'effet multiplicatif sur les

12. Cette notion de *logit* a déjà présenté au Chapitre 2.2.5.

odds [13] lorsqu'on incrémente x_i de 1, les valeurs des autres x_j (où $j \neq i$) étant gardées constantes.

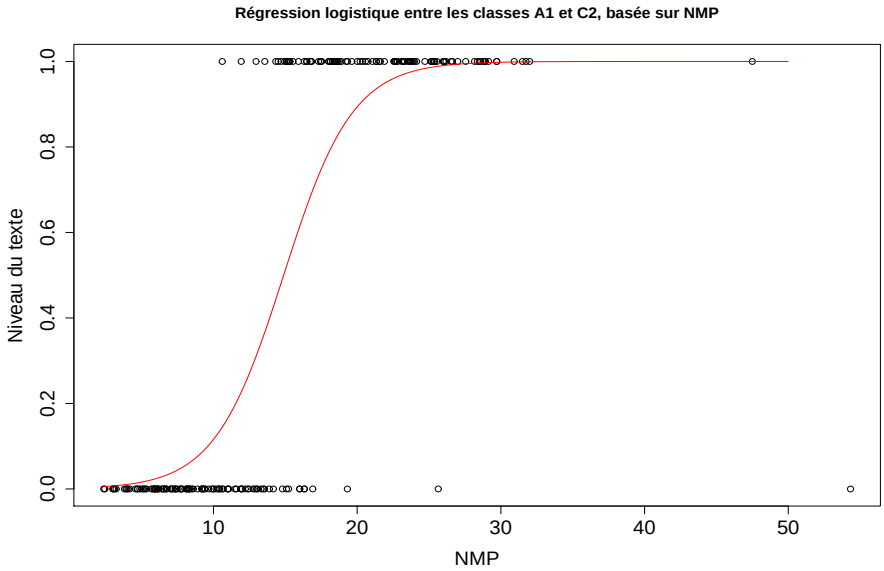


*Figure 5.9 – Exemple de fonction pour $E(Y = 1 \mid X = x)$, où X est la variable **NMP**. $Y = 0$ correspond aux textes de niveau A1, tandis que $Y = 1$ représentent ceux du niveau C2.*

La régression logistique multinomiale La RLM constitue une généralisation de cette technique pour un problème à K classes et se réalise à l'aide d'un modèle constitué de $K - 1$ sigmoïdes. Chacune d'elles compare la classe k à une catégorie de référence (souvent la première), afin de retomber sur le cas binaire. Ainsi, pour chacune de ces paires de classes (Y_k, Y_1), il existe une fonction décrite sous forme matricielle par l'équation suivante (Agresti, 2002, 268) :

$$log\frac{P(Y = k \mid X = \mathbf{x})}{P(Y = 1 \mid X = \mathbf{x})} = \alpha_k + \boldsymbol{\beta}_k^T \mathbf{x} \qquad k = 2, \cdots, K \tag{5.9}$$

où $\mathbf{x}$ est un vecteur observation, α_k et $\boldsymbol{\beta}_k^T$ représentent les paramètres du modèle pour la classe k. Sur la base de ces $K - 1$ équations, il est possible de calculer la probabilité qu'un texte appartienne au niveau de difficulté k pour un vecteur $\mathbf{x}$ donné. Cette probabilité est donnée par l'équation suivante [14] (Agresti, 2002, 271) :

$$P(Y = k \mid X = \mathbf{x}) = \frac{\exp(\alpha_k + \boldsymbol{\beta}_k^T \mathbf{x})}{1 + \sum_{k=2}^{K} \exp(\alpha_k + \boldsymbol{\beta}_k^T \mathbf{x})} \tag{5.10}$$

13. Il s'agit dans ce cas du rapport entre $\frac{p(x_i+1)}{p(x_i)}$.

14. Notons que pour la catégorie de référence ($k = 1$), α_1 et $\boldsymbol{\beta}_1^T = 0$. Aussi, lors du calcul de la probabilité qu'un texte appartienne à cette catégorie de référence, le numérateur vaut $\exp(0) = 1$.

L'estimation des paramètres d'un tel modèle s'effectue par maximum de vraisemblance à l'aide d'une procédure décrite dans Agresti (2002, 192).

La régression logistique ordinale Le défaut de la RLM est qu'elle repose sur $K-1$ équations, ce qui nécessite d'estimer $K-1*p+1$ paramètres. On obtient dès lors un modèle fort éloigné de l'idéal de parcimonie. La RLO ou modèle des chances proportionnelles (McCullagh, 1980) apporte une solution à ce problème, pour peu que la variable réponse soit ordinale. En effet, elle présente la particularité, par rapport au modèle multinational, d'attribuer une valeur unique aux coefficients β_k. Seule la valeur de l'ordonnée à l'origine (α_j) varie en fonction de la classe K considérée. Autrement dit, le modèle comporte aussi $K-1$ courbes, mais qui ont toutes la même forme et sont parallèles [15]. Il s'agit bien sûr d'une approximation qui ne peut, dans certains cas, pas être supportée par les données [16].

Pour une variable réponse ordinale Y à K niveaux et une matrice $\mathbf{x}$ constituée de N vecteurs observations à P variables explicatives, ce modèle peut s'écrire sous la forme suivante (Agresti, 2002, p. 275) :

$$\text{logit}[P(Y \leq k \mid \mathbf{x})] = \alpha_k + \boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{x} \tag{5.11}$$

Dans cette formule, la quantité d'intérêt correspond au logit de la probabilité que Y prenne une valeur égale ou inférieure à k étant donné $\mathbf{x}$. Grâce à la transformation logit, celle-ci peut ensuite être modélisée selon une relation linéaire avec les variables explicatives.

On peut également considérer la quantité inverse : la probabilité que Y prenne une valeur égale ou supérieure à k étant donné $\mathbf{x}$. La formule devient alors :

$$\text{logit}[P(Y \geq k \mid \mathbf{x})] = \alpha_k - \boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{x} \tag{5.12}$$

Ce logit cumulatif peut être développé de la manière suivante, ce qui le rend plus compréhensible :

$$\text{logit}[P(Y \geq k \mid \mathbf{x})] = \log \frac{P(Y \geq k \mid \mathbf{x})}{1 - P(Y \geq k \mid \mathbf{x})} = \log \frac{P(Y \geq k \mid \mathbf{x})}{P(Y < k \mid \mathbf{x})} \tag{5.13}$$

La troisième expression de l'égalité précédente met en évidence le lien entre la probabilité calculée dans le modèle des chances proportionnelles et le concept de chance (ou « odd »), qui a donné son nom à ce modèle.

Une fois un tel modèle entraîné par maximum de vraisemblance, l'estimation de la probabi-

15. De cette particularité provient le fait que l'on nomme également la RLO « modèle des pentes parallèles ».

16. Cette hypothèse de l'ordinalité des données nécessite d'être vérifiée, ce qui peut être fait à l'aide d'une forme chi-carré du « score test » (Agresti, 2002). Plus la valeur du chi-carré est petite, mieux le modèle est ajusté aux données.

lité qu'un texte D, représenté par le vecteur $\mathbf{d}$, appartienne à la classe k est calculée comme suit, quand $2 \leq k \leq K$:

$$P(Y = k \mid \mathbf{d}) = \text{logit}[P(Y \leq k \mid \mathbf{d})] - \text{logit}[P(Y \leq k-1 \mid \mathbf{d})] \tag{5.14}$$

Quand $k = 1$, $P(Y = 1 \mid \mathbf{x})$ est égal à $P(Y \leq k \mid \mathbf{x})$.

Pour en conclure avec les modèles logistiques, il convient encore d'expliquer comment y est opérée la sélection de variables. Nous l'avons réalisée en deux temps, suivant les indications de Hosmer et Lemeshow (1989). D'une part, seuls les prédicteurs ayant obtenu une corrélation dépassant un seuil δ (par exemple, $0,30$) au cours de l'analyse bivariée sont initialement considérés. Ce seuil ne doit toutefois pas être trop élevé, sous peine de négliger des facteurs susceptibles de devenir plus discriminants en association avec d'autres. Dans un second temps, un algorithme de sélection « pas-à-pas » est appliqué à cet ensemble de variables. Il existe deux versions de celui-ci. L'approche ascendante (« forward ») part d'un modèle nul (sans variable) et, itérativement, ajoute à chaque pas celle qui maximise un critère φ, jusqu'au moment où ce dernier n'augmente plus. À l'inverse, l'approche descendante (« backward ») part du modèle complet et retire itérativement la variable qui améliore le plus ce même critère φ.

Il apparaît donc que l'efficacité de la méthode dépend essentiellement de φ. On emploie communément pour celui-ci soit l'Akaike Information Criterion (AIC) de Akaike (1973) ou le Bayesian Information Criterion (BIC) de Schwarz (1978). Entre ces deux possibilités, Lebarbier et Mary-Huard (2004, 11) notent que le BIC permet plutôt de choisir le modèle le plus proche du phénomène à expliquer (utile pour une démarche explicative), alors qu'AIC est plus intéressant pour sélectionner un modèle efficace (démarche prédictive). C'est donc ce dernier qui a été retenu pour cette étude. Il s'exprime sous la forme suivante (Akaike, 1973) :

$$\text{AIC} = -2loglik + 2u \tag{5.15}$$

où u est le nombre de paramètres dans le modèle [17] et *loglik* correspond à la log-vraisemblance du modèle (cf. Hosmer et Lemeshow, 1989, 27). AIC vise donc à trouver un point d'équilibre entre la quantité de prédicteurs, dont dépend p, et l'efficacité du modèle à prédire les données.

17. Soulignons la différence entre u, le nombre de paramètres à estimer dans le modèle, et p, le nombre de prédicteurs. Ainsi, dans la RLM, la présence de $K-1$ équations fait que $u = K - 1 * p + 1$. Dès lors, supprimer une variable importe davantage que dans la RLO, où $u = K - 1 + p$.

Les arbres de décision

En dehors des modèles de régression, d'autres algorithmes ont été expérimentés. Il s'agit de classifieurs, qui font cette fois l'hypothèse que la variable réponse est nominale. Le plus simple de ceux que nous avons testés est un arbre de décision (Breiman *et al.*, 1984) dont le principe est de diviser l'espace des données en rectangles au moyen d'hyperplans qui représentent chacun un test effectué sur une variable (du type : X est plus grand que x). Chacun de ces sous-espaces contient ensuite un modèle extrêmement simple : il s'agit généralement d'une constante correspondant à la classe majoritaire dans cet espace.

La construction d'un arbre revient dès lors à définir une suite de segmentations successives de l'espace des données selon la procédure suivante : à chaque nœud, on sélectionne la variable qui divise le mieux les données en nœud-fils les plus purs. De manière plus formelle, l'impureté d'un noeud se calcule à l'aide de mesures telles que l'entropie, l'indice de Gini ou le taux d'erreurs de classification [18] et le choix du meilleur test est opéré en sélectionnant celui qui maximise le gain Δ de pureté, exprimé comme suit (Tan *et al.*, 2005, 160) :

$$\Delta = I(parent) - \sum_{j=1}^{J} \frac{N(v_j)}{N} I(v_j) \tag{5.16}$$

où N est le nombre d'observations dans le nœud parent, tandis que $N(v_j)$ correspond au nombre d'observations associées au nœud fils v_j et J, au nombre de niveaux de la variable explicative. Quant à la fonction d'impureté $I(\cdot)$, l'implémentation utilisée étant le module **rpart** de R (Therneau et Atkinson, 1997), il s'agit dans notre cas de l'entropie, laquelle se calcule comme suit pour un nœud v donné (Tan *et al.*, 2005, 158) :

$$Entropy(v) = -\sum_{k=0}^{K-1} p(k \mid v) \log_2 p(k \mid v) \tag{5.17}$$

où K représente toujours le nombre de niveaux de difficulté. L'objectif est de répéter cette démarche jusqu'à obtenir des nœuds terminaux ou feuilles les plus purs possible, c'est-à-dire ne comprenant idéalement que des observations appartenant à une même classe.

Ce modèle comporte plusieurs atouts [19]. Tout d'abord, cette technique est non-paramétrique : elle n'impose pas de conditions sur les distributions des prédicteurs et peut modéliser des relations non linéaires, voire non monotones, avec la variable réponse. De plus, elle s'adapte bien à un grand nombre de variables explicatives, étant donné que la procédure de sélection

18. Pour le détail de ces mesures, consulter Hastie *et al.* (2001, 271) ou, pour une approche plus abordable, Tan *et al.* (2005, 158-159).

19. Tufféry (2007, 250) présente une liste explicite des avantages et des défauts des arbres de décision dont nous nous sommes inspirés ici. Le lecteur désirant plus de détails à ce sujet peut s'y reporter.

de variables est cette fois intégrée à l'algorithme. Elle peut même détecter les interactions entre facteurs et reste peu sensible aux « outliers ».

Par contre, les arbres de décision seraient fort calculatoires et surtout efficaces pour des échantillons de grande taille (Besse, 2005, 71), ce qui n'est pas notre cas. Ils demandent également de fixer une série de paramètres destinés à élaguer l'arbre une fois construit. En effet, si celui-ci est trop touffu, il excellera sur les données d'entraînement, mais ne sera que très peu généralisable suite à un surapprentissage. À l'inverse, s'il est trop petit, il manquera de discernement. Il s'agit donc de naviguer entre ces deux extrêmes durant de la procédure d'élagage, laquelle est dirigée par les deux paramètres suivants (cf. Therneau et Atkinson, 1997, 12) :

– **minsplit** : spécifie le nombre minimum de données que doit contenir un noeud pour pouvoir encore être divisé ;
– **cp** : interdit toute division d'un nœud qui ne diminue pas le manque d'ajustement du modèle aux données par un facteur *cp*.

Cette procédure de sélection des paramètres optimaux étant propre à chaque modèle, elle est détaillée par la suite, dans les Sections 5.4.4 et 5.4.5.

Le bagging et le boosting

Au-delà des arbres de décision, il existe des méthodes plus complexes permettant de dépasser certaines de leurs limites : les techniques par agrégation de modèles. En effet, deux faiblesses supplémentaires des arbres de décision nous sont apparues au cours d'expériences préparatoires. D'une part, même lorsque les valeurs optimales de *cp* et *minsplit* étaient utilisées, il arrivait que l'une ou l'autre classe à prédire soit totalement absente du modèle, suite à un élagage trop important. D'autre part, le choix des prédicteurs et leur ordonnancement se sont révélés fort dépendants des données d'apprentissage et sujets à une grande variabilité. Or, Tufféry (2007, 251) explique qu'un arbre de décision ne peut mettre à jour que des optimums locaux et non globaux. Une fois une variable retenue comme critère de division d'un noeud, ce choix n'est en effet plus remis en question. Par conséquent, si pour un individu qui a toutes les caractéristiques d'un groupe précis sauf en une variable, le test s'applique d'abord sur cette variable, l'individu risque alors d'être mal classé.

C'est cette variabilité au niveau des classes représentées et de l'ordonnancement des tests que les approches par agrégations de modèles visent à réduire. Leur principe consiste à entraîner plusieurs modèles sur M corpus différents obtenus à partir d'un seul. Ils diffèrent toutefois dans la méthode mise en pratique cela. Le premier d'entre eux, le bagging ou « bootstrap aggregating » (Breiman, 1996), crée à chaque itération un sous-corpus par ré-

échantillonnage du corpus de base, qui permet d'entraîner un nouveau modèle. Par la suite, le niveau d'un texte peut être défini comme la classe majoritaire lors du vote de ces M classifieurs. Cette procédure comporte l'avantage de réduire la variance d'un modèle et s'applique donc particulièrement bien à des classifieurs instables tels que les arbres de décision.

Le boosting, mis au point par Freund et Schapire (1996), conserve par contre le même corpus tout au long du processus. Il présente la particularité d'être adaptatif, car il se concentre sur les données difficiles à modéliser. Il en existe plusieurs variantes, mais l'algorithme utilisé dans cette étude est *AdaBoost* (Freund et Schapire, 1996), disponible dans le module *ada* de R. Celui-ci opère selon une procédure itérative à M passes qui débute par l'attribution d'un poids $w_i = \frac{1}{N}$ à chacune des N observations du corpus d'entraînement. Le classifieur $h_m(\mathbf{x})$ ($1 \leq m \leq M$) est ensuite entraîné sur ce jeu de données et une estimation de son taux d'erreur ϵ_m est obtenue [20]. Sur la base de cette estimation de l'erreur, on peut alors recalculer le poids α_m du classifieur, avant de réévaluer les poids w_i de chacune des observations, donnant plus d'importance aux données mal classées et moins à celles qui l'ont été correctement. À la fin de cette procédure, on obtient un ensemble de M classifieurs dont la prédiction est obtenue via un vote pondéré de la façon suivante (Meir et Rätsch, 2003, 118) :

$$f_{Ens}(\mathbf{x}) = \sum_{m=1}^{M} \alpha_m h_m(\mathbf{x}) \tag{5.18}$$

L'une des forces principale du boosting est qu'il atteint des performances élevées en combinant un ensemble de classifieurs *faibles*, c'est-à-dire classant à peine mieux que le hasard (Meir et Rätsch, 2003, 118-119). Dans notre cas, deux types d'arbres ont été utilisés comme classifieurs *faibles* : un arbre de décision binaire (**stump**) et un arbre de décision à quatre feuilles (**four**). L'emploi de structures aussi simples [21] pour un problème à K classes s'explique par le fait que le boosting ne convient normalement qu'à une variable réponse de type binaire. Nous avons donc dû adapter cette technique pour un problème à K classes en recourant à la stratégie du « un contre tous » (Culp *et al.*, 2006, 22-24). Celle-ci développe K ensembles de modèles, en opposant chaque classe k au reste des données. On obtient donc K prédicteurs qui, pour un texte donné, vont chacun estimer la probabilité que ce texte appartienne à la classe k ou pas. C'est pour cette raison qu'il est possible d'obtenir de bonnes performances avec des arbres binaires ou quaternaires. Une fois ces K modèles obtenus, reste alors à attribuer à chaque texte la classe pour laquelle $f_{Ens,k}(\mathbf{x})$ est la plus élevée.

Soulignons enfin qu'existent aussi les « random forests » de Breiman (2001). Elles amé-

20. Pour le détail des formules utilisées pour calculer ce taux d'erreur et repondérer les observations, se reporter à Meir et Rätsch (2003).

21. Par le terme « simple », nous voulons signifier que ces arbres ne comportant que deux ou quatre feuilles, ils ne peuvent classer qu'un nombre équivalent de classes, inférieurs à K, dans notre cas.

liorent le bagging en étendant le rééchantillonnage aux prédicteurs. Dès lors, à chaque division de noeud, seul un sous-ensemble de prédicteurs est considéré lors de sélection du meilleur attribut. Nous ne les avons toutefois pas retenues dans cette étude, car leurs performances étaient sensiblement similaires à celles du boosting lors de nos expériences préliminaires.

Les machines à vecteurs de supports

La dernière méthode envisagée est un SVM. Une version linéaire de cet algorithme existait depuis 1963, mais son succès actuel découle de ce que Boser *et al.* (1992) en ont proposé une version applicable à des problèmes binaires non linéaires, grâce au « kernel trick ». Cette technique opère en deux étapes : (1) l'espace d'origine est transformé en un espace ϕ de plus grande dimension, de sorte qu'il devienne possible de séparer linéairement les données transformées $\phi(\mathbf{x})$; (2) un hyperplan est ensuite utilisé comme séparateur linéaire entre les deux classes. Celui-ci est exprimé selon l'équation (Tufféry, 2007, 357) :

$$f(x) = \mathbf{w}^T\mathbf{x} + b \tag{5.19}$$

Une fois les paramètres $\mathbf{w}$ et b connus, il est possible de classer un nouveau vecteur d'observation $\mathbf{z}$ comme suit (Tan *et al.*, 2005, 260) :

$$y = \begin{cases} 1, & \text{si } \mathbf{w}^T\mathbf{z} + b > 0 \\ -1, & \text{si } \mathbf{w}^T\mathbf{z} + b < 0 \end{cases}$$

La particularité des SVM, que l'on appelle aussi « séparateurs à vaste marge », est de chercher à maximiser la marge située autour de cet hyperplan afin de réduire les erreurs lors de la généralisation du modèle. Concrètement, cette opération s'effectue en remplaçant cet hyperplan par deux autres, correspondant à chacune de ses deux marges, et à maximiser la distance entre ces deux plans (voir Figure 5.10). Dans un cas non linéaire et non séparable tel que le nôtre, la procédure de maximisation de la marge et d'estimation des paramètres est relativement complexe, mais il s'agit néanmoins d'une situation bien documentée par ailleurs (Hastie *et al.*, 2001; Tan *et al.*, 2005; Tufféry, 2007), que nous ne développerons donc pas ici.

Lorsqu'on entraîne un modèle SVM, il importe toutefois d'optimiser un certain nombre de paramètres. Il existe apparemment certaines méthodes logiques pour découvrir les valeurs idéales, mais la majorité des chercheurs agissent par recherche empirique, en testant en validation croisée un grand nombre de valeurs au moyen d'une « grid search », comme cela

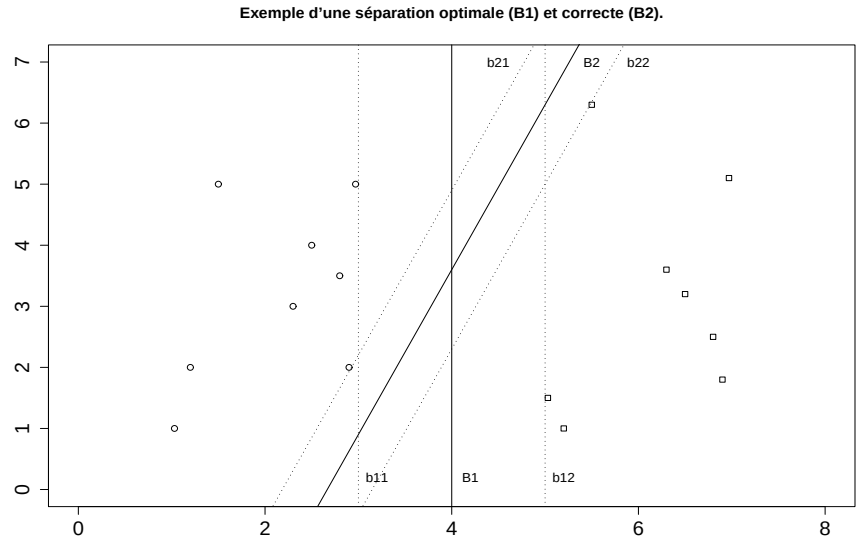


Figure 5.10 – Exemple de deux hyperplans séparateurs. Le premier B1, maximise la marge comprise entre b11 et b12, tandis que B2 sépare correctement les données, mais possède une marge plus réduite.

a déjà été fait pour la recherche du k à la Section 5.2.6. Ces différents paramètres sont :

– **kernel** : le choix du kernel influence le type de problèmes qui peuvent être traités. Sur les indications de Hsu *et al.* (2009, 4-5), nous avons opté pour le Gaussian Radial Basis Function (RBF), qui peut traiter des problèmes non linéaires et permet de limiter le nombre d'hyperparamètres à deux : C et γ. Ce dernier aspect est important, car la recherche des meilleures valeurs (C, γ) doit s'effectuer de manière conjointe, ce qui augmente considérablement le nombre d'essais requis.
– **le coût (**C**)** : représente la pénalité attribuée en cas de mauvaise classification d'une observation lors de l'entraînement. Plus celle-ci est élevée, plus le modèle utilise des marges rigides et ajustées aux données, au risque de mal se généraliser sur de nouvelles observations. Par contre, si C est trop bas, le modèle tolère un grand nombre d'erreurs d'entraînement et risque de définir des marges trop simples par rapport au problème.
– γ : représente le paramètre σ dans la fonction du RBF. Sa valeur conseillée par défaut est égale à $\frac{1}{N}$, où N est le nombre de données d'entraînement.

Enfin, indiquons que, comme le boosting, l'approche par SVM nécessite d'être adaptée pour s'appliquer à un problème à K classes. Cette fois, nous avons utilisé la technique du « un contre un » pour ce faire, qui diffère de la « un contre tous » employée pour le boosting en ce sens qu'elle consiste à entraîner C_K^2 classifieurs binaires (ici, 15 ou 36 selon que K vaut 6 ou 9) qui votent chacun pour une classe plutôt qu'une autre. La classe majoritaire au final est retenue.

Conclusion

Cette section a présenté six modèles statistiques qui permettent de répondre, chacun avec leurs forces et leurs faiblesses, à la question de la pondération des variables via une approche automatisée. Comme le lecteur a pu s'en rendre compte, les algorithmes employés, bien que récents et performants, sont employés dans leur version classique. Notre objectif n'est en effet pas de rechercher une hausse des performances via une optimisation complexe des modèles. D'une part, cela reviendrait à trop s'écarter du champ de la linguistique appliquée dans lequel se situe principalement cette étude ; d'autre part, les quelques expérimentations supplémentaires effectuées durant la phase préparatoire de cette étude n'ont pas conduit à des améliorations significatives. Par conséquent, ce sont ces six modèles qui seront utilisés pour l'entraînement de notre formule.

Une fois réglée la question du choix des variables et de leur pondération au sein des différents modèles, il ne reste plus qu'à évaluer les performances de ceux-ci sur un corpus d'évaluation, afin de retenir celui qui classe le mieux les textes de notre corpus. Une procédure que décrit la section suivante de ce chapitre.

5.4 Comparaison des modèles

5.4.1 Introduction

Cette dernière section, qui clôt notre travail, s'attelle à l'entraînement et l'évaluation de modèles combinant plusieurs des variables étudiées précédemment. Elle débute par deux précisions méthodologiques supplémentaires. D'une part, la nécessité d'estimer certains hyperparamètres au sein des divers modèles statistiques a rendu indispensable l'implémentation d'une double procédure de validation croisée, détaillée à la Section 5.4.2, et qui sous-tend la création de deux corpus différents : l'un de développement, l'autre de test. D'autre part, la Section 5.4.3 décrit une série de mesures d'évaluation apte à juger de l'efficacité des formules de lisibilité obtenues à l'aide des différents algorithmes statistiques. Elle présente aussi un test statistique permettant de conclure à la supériorité d'un modèle sur un autre.

Par la suite, une synthèse des nombreuses expérimentations réalisées est proposée sous deux sections différentes. La première (*cf.* Section 5.4.4) compare les modèles résultant de la sélection manuelle de variables, tandis que la seconde (*cf.* Section 5.4.5) discute les modèles obtenus via une procédure automatisée de sélection des variables. Enfin, un dernier point (*cf.* Section 5.4.6) analyse en détail la formule retenue comme la plus adaptée à l'évaluation de la lisibilité de textes de FLE, la comparant notamment avec les autres réalisations en lisibilité.

5.4.2 Méthodologie d'évaluation : les corpus

Il ressort de la discussion sur les modèles statistiques que les échantillons assemblés dans la première partie de ce chapitre ne peuvent être réutilisés tels quels. En effet, ils comportent l'ensemble des 108 textes disponibles par niveau (sauf pour la LSA), ce qui ne laisse aucun document C2 de libre pour évaluer les différents hyperparamètres des modèles statistiques sur des données différentes que celles servant à évaluer les modèles eux-mêmes. Or, l'usage d'une procédure en validation croisée[22] est requis pour éviter que les performances des modèles ne soient surévaluées suite à un choix ad hoc et peu généralisable de ces paramètres (Duda *et al.*, 2001, 483). Il a donc été nécessaire de diviser, une fois pour toutes, l'ensemble des textes disponibles en deux échantillons. Le premier, qui comporte 40 textes par niveau, a servi de corpus de développement, tandis que le second, comprenant cette fois 68 textes par niveau, a été utilisé pour évaluer les modèles.

Optimisation des méta-paramètres

Concrètement, les *Corpus6CE* et *Corpus9CE* ont subi une première phase d'échantillonnage, au cours de laquelle 40 textes ont été sélectionnés aléatoirement par niveau. Il en est résulté *Dev6CE* et *Dev9CE*, qui comportent respectivement 240 et 360 observations. La procédure de suppression automatique des données aberrantes décrites à la section 5.2.3 leur a ensuite été appliquée, avant qu'ils ne servent à déterminer les valeurs optimales des différents hyperparamètres, ainsi qu'à effectuer la sélection automatisée des variables. Pour ce faire, nous avons recourru à une « grid search » où, à chaque valeur ($iter$, $nbagg$) ou paire de valeurs (C, γ et cp, $minsplit$), correspond un résultat de la fonction objective utilisée, à savoir le R ou l'exactitude contiguë du modèle. Notons que ce résultat a été obtenu via un processus de validation croisée à dix échantillons, afin de rendre son estimation plus précise.

Évaluation des performances des modèles

Dans un second temps, un nouvel échantillonnage a été effectué sur le reste des textes et a conduit à la création de *Test6CE* et *Test9CE*, qui comportent respectivement 408 et 612 textes avant suppression des « outliers ». C'est sur ces deux jeux de données que les modèles statistiques ont été entraînés et évalués à l'aide d'une validation croisée à dix échantillons.

Par ailleurs, rappelons qu'il a aussi été nécessaire de constituer un nouvel espace vectoriel pour le calcul des variables basées sur la LSA, étant donné que certains des textes de l'échantillon défini précédemment (*cf.* Section 5.2.6) risquaient de se retrouver dans *Test6CE* ou

22. Signalons que la procédure utilisée est plus exactement une procédure de validation croisée stratifiée, où la proportion des observations dans les différentes classes reste relativement stable.

Test9CE. Pour ce faire, nous avons, comme précédemment, employé *Dev6CE* ou *Dev9CE*, complétés par l'ensemble des textes non repris dans *Test6CE* ou *Test9CE*. Le nombre de dimensions conservées après la SVD dans la LSA est toujours de 15, comme dans la première partie.

5.4.3 Méthodologie d'évaluation : les mesures

Une fois rassemblées les données de développement et de test, il reste une seconde question méthodologique à résoudre : comment évaluer l'efficacité d'un modèle et comment déterminer la supériorité effective d'un modèle sur un autre, sachant qu'une légère différence de performances peut n'être due qu'au hasard ?

En ce qui concerne l'évaluation d'un modèle donné, plusieurs mesures d'évaluation ont été proposées dans la littérature. Parmi celles-ci, la plus commune est certainement R, le coefficient de corrélation multiple, largement employé dans la lisibilité classique et structuro-cognitive (*cf.* Chapitre 2). Comme nous l'avons déjà expliqué (*cf.* Chapitre 2.2.2), elle correspond simplement à la corrélation de Pearson calculée entre les niveaux réels des textes (Y) et les niveaux prédits par le modèle ($\hat{Y}$). Toutefois, elle se calcule généralement plutôt comme suit (Faraway, 2002, 21-22) :

$$R = \sqrt{1 - \frac{SC_{erreur}}{SC_{total}}} = \sqrt{1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y}_i)^2}} \tag{5.20}$$

À côté de cette mesure bien connue, d'autres, qui sont davantage adaptées à un problème de classification, sont récemment apparues au sein de la lisibilité computationnelle. Il s'agit de l'exactitude et de l'exactitude contiguë. L'exactitude (acc) ou « accuracy » en anglais, est simple à calculer : il s'agit du nombre de prédictions correctes sur le nombre total d'observations. Formellement, elle est définie sur la base de la matrice de dispersion, qui est en fait une table de contingence présentant la distribution bivariée des classes d'origine par rapport aux prédictions du modèle. Sa structure est représentée dans le Tableau 5.27 pour le cas binaire.

		Prédictions	
		classe = 1	classe = 0
Classes réelles	classe = 1	f_{11}	f_{10}
	classe = 0	f_{01}	f_{00}

Tableau 5.27 – Exemple de matrice de confusion pour un problème de classification binaire. Chaque entrée f_{ij} correspond au nombre d'observations de la classe i prédits comme appartenant à la classe j.

À partir de cette matrice de confusion, que nous utiliserons également pour analyser en détail la répartition des prédictions de certains modèles, il est possible de formaliser le calcul de l'exactitude comme suit (Tan *et al.*, 2005, 149) :

$$\text{Exactitude} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}} \tag{5.21}$$

L'exactitude contiguë ($acc - cont$), ou « adjacent accuracy » en anglais, est quant à elle définie par Heilman *et al.* (2008) comme la proportion de prédictions qui s'éloignent au plus d'un niveau par rapport à celui qui a été attribué au texte par un humain. Son emploi se justifie par la difficulté qu'ont divers experts humains à accorder leur jugement sur un ensemble de textes, comme nous l'avons observé au sein des manuels (*cf.* Chapitre 4). Il faut toutefois signaler qu'elle fournit des valeurs particulièrement optimistes lorsqu'il y a peu de catégories (elle vaut 100% dans le cas binaire) et qu'elle devrait toujours être comparée à l'exactitude contiguë obtenue en utilisant une classification au hasard.

Ces deux mesures s'ajoutent donc à l'appareil méthodologique utilisé pour évaluer la qualité respective des modèles entraînés. Cependant, ces trois outils (R, acc et $acc-cont$) présentent tous un défaut. L'acc et l'$acc-cont$ donnent certes une idée précise du pourcentage de textes correctement ou presque correctement classifiés, mais n'apportent pas d'informations sur le degré de gravité des erreurs. Ainsi, un modèle qui ne se trompe jamais de plus de deux niveaux ne sera pas nécessairement moins pénalisé qu'un second produisant des erreurs de quatre à cinq niveaux, telles que d'attribuer le niveau C2 à un texte A1. À l'inverse, R prend ce phénomène en compte, mais produit un résultat qui n'est pas facilement interprétable.

Dès lors, Heilman *et al.* (2007, 465) défendent l'emploi de l'erreur quadratique moyenne (MSE) comme mesure d'évaluation à la place de l'exactitude. En effet, ils estiment que prédire la lisibilité de textes s'assimile davantage à une tâche de régression que de classification. Rappelons que la MSE correspond à la moyenne arithmétique des carrés des écarts entre les prévisions et les classes de référence (Witten et Frank, 2005, 178) :

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \tag{5.22}$$

Le fait de travailler avec le carré des erreurs rend la MSE capable d'attribuer une importance variable aux erreurs de prédiction en fonction de leur magnitude : elle diminue les petits écarts (inférieurs à 1) et accentue les grands. Cependant, elle utilise également une unité peu parlante, à savoir le carré des niveaux. C'est pourquoi nous avons préféré recourir à erreur

type (RMSE), qui correspond à la racine carrée de la MSE (Witten et Frank, 2005, 178) :

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{5.23}$$

Cette quatrième mesure retenue pour l'évaluation des modèles est sujette aux inconvénients de sa qualité principale : la mise au carré des erreurs permet certes de pénaliser les erreurs graves, au risque de surévaluer celles-ci par rapport à l'ensemble des prédictions dans le cas des « outliers ». Comme il est n'est pas exclu que certains d'entre eux passent outre la procédure de suppression automatique mise en place, nous avons également employé une cinquième mesure : l'erreur moyenne absolue (MAE). Celle-ci se calcule comme suit (Witten et Frank, 2005, 178) :

$$MAE = \frac{1}{N}\sum_{i=1}^{N} \mid \hat{y}_i - y_i \mid \tag{5.24}$$

Comme l'erreur type, la MAE présente l'avantage d'être exprimée dans l'unité de la variable réponse. Dans notre cas, elle peut s'interpréter comme le nombre de niveaux dont un classifieur se trompe en moyenne. Par contre, elle n'accentue pas la gravité des erreurs les plus lourdes.

En résumé, nous disposons donc de cinq mesures d'évaluations capables d'analyser la qualité d'un modèle : R, acc, $acc - cont$, RMSE et MAE. Comme l'estimation de ces mesures est obtenue via une procédure de validation croisée à dix échantillons, les valeurs rapportées représentent en réalité la moyenne obtenue sur les dix échantillons de test. Comme nous l'avons déjà expliqué précédemment, celle-ci constitue dès lors une estimation $\sqrt{10}$ fois plus précise que celle qu'on obtiendrait à partir d'un seul jeu de données. Mais, le recours à cette procédure permet également de calculer l'écart-type de ces 5 mesures, lequel constitue un bon indicateur de la stabilité du modèle sur de nouvelles données. Afin de ne pas surcharger les tableaux, celui-ci n'a cependant été rapporté que pour l'exactitude et l'exactitude contiguë.

Cependant, ces cinq mesures ne suffisent pas pour comparer les performances de deux modèles différents, ni pour déterminer s'il existe une différence significative entre ceux-ci. En effet, un score légèrement supérieur de l'un d'eux pourrait s'expliquer uniquement par un effet d'échantillonnage et pourrait être insuffisant pour généraliser la supériorité de ce modèle à l'ensemble des situations réelles.

Pour opérer ces comparaisons, il est possible de recourir au test de McNemar (McNemar, 1947) ou à un test-T pairé (Witten et Frank, 2005, 153-157) pour autant que l'évaluation des modèles se fasse via une procédure de validation croisée à dix échantillons où ces der-

niers sont similaires d'un modèle à l'autre. Or, c'est justement l'approche que nous avons implémentée. En effet, la validation croisée à dix échantillons fournit une estimation robuste des performances de nos modèles, tandis que le fait de garder les mêmes échantillons (via un échantillonnage systématique) permet de contrôler l'erreur d'échantillonnage. Par conséquent, il est apparu logique d'utiliser le test-T pairé pour discriminer nos modèles.

5.4.4 Comparaison des modèles experts

Introduction

La première série de modèles à évaluer découle d'une sélection de variables par expertise. Ceux-ci visent surtout à tester l'une des hypothèses majeures de notre thèse, à savoir que l'un des moyens de dépasser les performances des formules classiques consisterait à ajouter, aux variables lexico-syntaxiques classiques, un nombre réduit de prédicteurs traitant d'autres dimensions textuelles. Dans la première partie de ce chapitre, nous avons identifié plusieurs bons candidats à ce rôle, ce qui a permis de définir trois ensembles de variables, déjà présentés précédemment (*cf.* Section 5.3.2). L'objectif de cette section est d'expérimenter l'efficacité de chacun des six modèles statistiques sur ces trois sélections et d'éclaircir les questions de recherche suivantes :

– Quel est le modèle qui se comporte le mieux au sein de chaque sélection ?
– Est-ce que les meilleurs modèles basés sur *Expert1* et sur *Expert2*, dénommés respectevie-ment $bestE1$ et $bestE2$, qui prennent en compte quatre dimensions textuelles, obtiennent des performances supérieures à la « baseline » ?
– Est-ce que l'ajout de quatre prédicteurs rend $bestE2$ supérieur à $bestE1$?

Pour répondre à ces questions, il a été nécessaire d'effectuer une série de tests statistiques destinés à confronter les modèles. Afin d'en limiter le nombre, nous avons défini une série de comparaisons planifiées (*a priori*) (Howell, 2008, 359-360), qui évitent de recourir à des comparaisons *post hoc* (Howell, 2008, 359-360). En effet, nos analyses préalables (François, 2009b) ayant montré que ce sont les classifieurs par RLM et RLO qui se comportent le mieux, nous nous sommes généralement limité à les confronter avec les résultats des SVM, que nous n'avions pas encore expérimentées, pour chaque ensemble de variables. Par ailleurs, $bestE1$ et $bestE2$ ont été comparés à la « baseline » et entre eux. Cela fait donc trois tests par ensemble de variables et trois tests pour les comparaisons entre chaque paire d'ensembles [23]. Ces opérations ont été conduites sur les trois ensembles de variables définis

23. Nous n'avons dès lors pas considéré nécessaire de diminuer la valeur d'α, bien que cela soit l'usage, au moins dans le cas de tests *post hoc*. En effet, dans ce cas, le fait de travailler sur la base d'une validation croisée à dix échantillons n'aurait permis de détecter des différences significatives que dans les cas extrêmes, ce qui nous a semblé une position trop conservatrice. Toutefois, sachant que certains de nos lecteurs peuvent être

plus haut : la « baseline », qui vise à simuler les performances d'une formule classique sur nos données ; le modèle *Expert1* à quatre variables et le modèle *Expert2* à huit prédicteurs.

La baseline

Pour la création de la « baseline », deux variables classiques et simples à implémenter ont été utilisées : le nombre moyen de mots par phrase (**NMP**) et le nombre moyen de lettres par mot (**NLM**). Comme nous l'avons déjà souligné, cette formule simple vise à donner une idée des performances qu'une formule de type classique pourrait atteindre pour notre problème, sachant qu'il n'en existe pas qui puisse être directement utilisée pour prédire la difficulté de textes selon l'échelle du CECR.

Avant d'évaluer ses performances en validation croisée, il a été nécessaire de déterminer les valeurs d'une série de méta-paramètres pour certains des modèles statistiques. Ainsi, les arbres de décision requierent de définir cp et $minsplit$; le bagging demande d'identifier le nombre idéal d'échantillons bootstrap ; le boosting laisse le choix entre les arbres binaires ou quaternaires et le nombre d'itérations ; enfin, les SVM nécessitent de sélectionner la paire optimale de valeurs pour C et γ. Nous avons testé plusieurs valeurs pour chacun de ces paramètres à l'aid d'une « grid search » dont la fonction objective est tantôt R, tantôt $acc - cont$. Pour chaque valeur de paramètre, le résultat de cette fonction a été calculé via une validation croisée à 10 échantillons, afin de rendre son estimation plus robuste.

Parmi ces problèmes d'optimisation, les deux cas les plus simples sont le bagging et le boosting, car il s'agit simplement d'investiguer un continuum numérique à une dimension. Cette exploration supposant toutefois de tester un certain nombre de valeurs, nous avons recouru à des graphiques pour représenter le R ou l'$acc - cont$ du modèle correspond à chacune d'elles. Afin de ne pas surcharger ce chapitre, ceux-ci ont été reportés à l'annexe D.

Pour le bagging, les valeurs utilisées sont : 10, 15, 20, 25, 30, 50, 75, 100, 120, 150, 175, 200, 250, 500, 750 et $1\,000$. On note sur la Figure 5.11 pour les modèles à 6 classes [24], une grande variabilité des résultats pour les premières valeurs, avant qu'apparaisse une stabilisation des performances à partir de 150 itérations. C'est pourquoi nous avons retenu cette dernière valeur pour $nbagg$, même si elle ne constitue pas le maximum de la fonction. En effet, un nombre inférieur d'itérations peut certes conduire à un R légèrement supérieur dans certains cas, mais rien ne garanti que, sur les données de validation, il ne correspondera pas à l'un

en désaccord avec cette position, les p-valeurs précises ont été rapportées afin que ceux-ci puissent analyser les résultats à leur façon. Le seuil que nous avons utilisé pour l'ensemble des tests correspond à une erreur $\alpha = 0,05$.

24. Le graphique du modèle à 9 niveaux n'est pas reproduit ici, mais est disponible en annexe. On y remarque la même tendance.

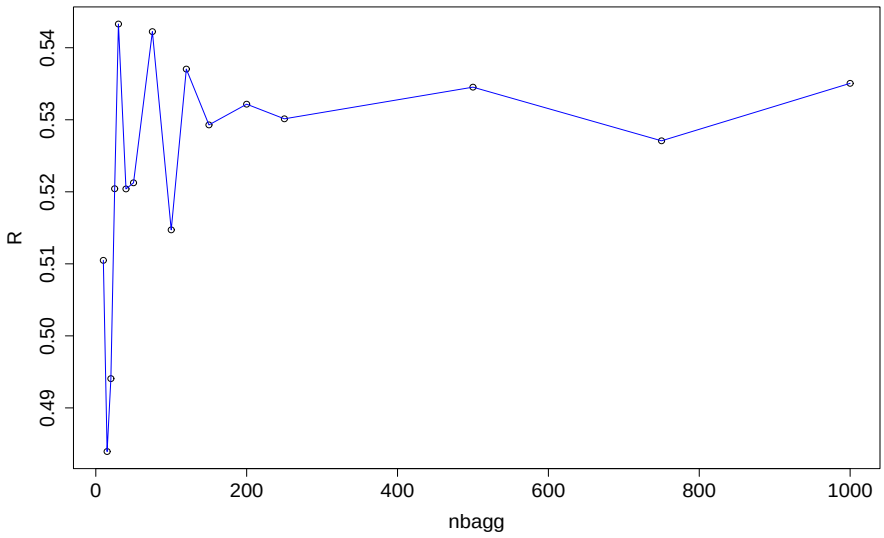


Figure 5.11 – Exemple de graphique présentant les résultats de la fonction objective (ici, R) en fonction du paramètre nbagg dans le bagging appliqué au Dev6CE.

des creux que l'on peut observer sur le graphique.

Quant au boosting, nous avons d'abord employé le même ensemble de valeurs que pour le bagging (*cf.* Annexe D), mais observant une tendance nette à la détérioration des performances à partir de 200 itérations sur *Dev6CE*, nous avons ensuite utilisé l'ensemble suivant : 10, 15, 20, 25, 30, 40, 50, 75, 100, 120, 150, 175, 200, 250, afin d'alléger les temps de calcul. En effet, l'entraînement d'un modèle à 1000 itérations est extrêmement calculatoire ! Sur la base de ces nouvelles valeurs et, malgré une variabilité importante des résultats en fonction des valeurs, il apparaît que R augmente légèrement jusqu'à 30 ou 40 itérations, pour ensuite diminer progressivement. Par ailleurs, les arbres quaternaires obtiennent des scores légèrement meilleurs que les arbres binaires. Par conséquent, les paramètres retenus pour les deux corpus sont : arbres quaternaires avec 40 itérations.

Plus complexe est le cas des arbres de décision. En effet, il s'agit cette fois d'explorer un espace à deux dimensions constitué par cp et $minsplit$. Nous basant sur nos travaux précédents, nous avons testé quatre valeurs pour $minsplit$ (5, 10, 15 et 20) et six pour cp ($0,1$; $0,05$; $0,01$; $0,005$; $0,001$; $0,0005$), ce qui donne vingt-quatre combinaisons possibles. Comme l'illustre le Tableau 5.28 pour *Dev6CE*, il est apparu, d'une part, que les valeurs cp de $0,1$ et $0,05$ sont trop grandes et produisent des arbres trop petits (qui comprennent 2 ou 3 noeuds) et, d'autre part, qu'il existe plusieurs combinaisons efficaces (en gras), qui engendrent des arbres de taille similaire (6 à 7 noeuds).

Afin d'éclairer un peu la question, nous avons testé à nouveau ces mêmes valeurs, mais prenant cette fois l'exactitude contiguë comme critère (*cf.* Tableau 5.28). Il en est ressorti

cp	$minsplit$	R	$acc-cont$
$0,1$	5	$0,592$	$55,2$
$0,05$	5	$0,595$	60
$0,01$	5	$0,618$	67
$0,005$	5	$0,619$	$67,8$
$0,001$	5	$\mathbf{0,62}$	**70**
$0,0005$	5	$0,595$	$66,1$
$0,1$	10	$0,592$	55
$0,05$	10	$0,60$	$56,5$
$0,01$	10	$0,618$	$68,69$
$0,005$	10	$\mathbf{0,624}$	70
$0,001$	10	$0,605$	$\mathbf{72,6}$
$0,0005$	10	$0,611$	66
$0,1$	15	$0,591$	$55,2$
$0,05$	15	$0,611$	57
$0,01$	15	$\mathbf{0,643}$	$68,3$
$0,005$	15	$0,623$	$66,1$
$0,001$	15	$0,612$	$68,3$
$0,0005$	15	$0,623$	$67,4$
$0,1$	20	$0,592$	$55,2$
$0,05$	20	$0,585$	$56,5$
$0,01$	20	$\mathbf{0,627}$	$69,1$
$0,005$	20	$0,626$	$64,8$
$0,001$	20	$0,627$	$67,8$
$0,0005$	20	$0,627$	$69,6$

Tableau 5.28 – Performances des modèles par arbre de décision en fonction des paramètres cp et $minsplit$, estimées par validation croisée à dix échantillons sur Dev6CE.

une zone plus favorable, située entre $0,001 \leq cp \leq 0,01 \mid minsplit = 10$, laquelle a été retenue. Nous y avons répété une recherche via une granularité plus petite, mettant à jour la combinaison $cp = 0,003$ et $minsplit = 10$[25]. Soulignons que la taille des arbres ainsi constitués ne dépasse pas 6 à 7 feuilles. Dès lors, le problème décrit précédemment à propos des arbres de décision, à savoir le fait que certains niveaux ne soient pas associés à une des feuilles, a été observé à nouveau. Pour *Dev9CE*, la même procédure en deux temps a été appliquée et a conduit à sélectionner les valeurs suivantes : $cp = 0,01$ et $minsplit = 5$.

Enfin, reste à effectuer une recherche similaire pour les paramètres des SVM : C et γ. Pour γ, rappelons qu'il est conseillé de prendre comme valeur $\frac{1}{N}$, où N est le nombre d'observations. Appliqué à nos corpus, cela donne $0,0042$ pour *Dev6CE* et $0,0027$ pour *Dev9CE*. Nous avons dès lors essayé des valeurs proches de cette recommandation et d'autres nettement plus éloignées, obtenant le vecteur suivant : $0,0001$; $0,0005$; $0,001$; $0,002$; $0,004$; $0,006$; $0,01$; $0,1$ et $0,5$. En ce qui concerne le coût, qui pénalise les points non séparables, fréquents dans nos jeux de données, il est probable qu'il ne doive pas être trop élevé. C'est pourquoi

25. Notons cependant que, pour cette seconde recherche, $acc - cont(cp = 0,003; minsplit = 10) = 69,6\%$, tandis que $acc - cont(cp = 0,001; minsplit = 10)$ est tombé à $68,7\%$. C'est la preuve que les résultats sont sujets à une variabilité importante, rendant ce processus imprécis. Rien ne permet donc d'affirmer que l'une de ces deux combinaisons est meilleure que l'autre.

les quinze valeurs suivantes ont été utilisées : 1, 2, 5, 8, 10, 12, 15, 20, 25, 40, 50, 75, 100, 500 et 1000. Cela donne pas moins de $15 * 9$, à savoir 141 modèles à entraîner.

Alors que nous avions retenu les valeurs de $\gamma = 0,008$ et $C = 7$ pour des expériences précédentes sur les modèles à 6 niveaux, cette recherche plus large a mis en évidence l'intérêt de valeurs de γ plus élevées, telles que $0,1$, voire même $0,5$. En ce qui concerne C, trois valeurs sont apparues intéressantes : 5, 20 et 25. Après la seconde passe de la recherche, nous avons retenu la paire $\gamma = 0,05$ et $C = 25$, qui atteint une exactitude contiguë de $70,9\%$, pour le modèle à 6 niveaux et la paire $\gamma = 0,01$ et $C = 40$ pour le modèle à 9 niveaux.

Une fois cette étape terminée (les valeurs des paramètres retenus sont présentés dans le Tableau 5.29), il a été possible de calculer, sur les échantillons de tests, les performances des modèles qui en découlent. Leurs résultats sont également repris au Tableau 5.29. On y observe tout d'abord que ce ne sont pas les mêmes algorithmes qui se démarquent selon qu'on considère l'exactitude ou des mesures de la dispersion des erreurs. Ainsi, RLM obtient la meilleure exactitude sur *Test6CE,* mais se comporte moins bien en $acc-cont$ que les SVM ou le boosting. Ce dernier constitue d'ailleurs une bonne surprise, car il apparaît comme le modèle le plus constant en termes d'$acc-cont$ sur les deux échantillons de test, alors que nos expériences précédentes François (2009b) ne l'avaient pas détecté comme particulièrement efficace à ce niveau.

Modèle	Paramètres	Corpus	R	acc	$acc-cont$	$rmse$	mae
Hasard	/	Test6CE	/	$16,6\%$	$44,4\%$	/	/
		Test9CE	/	$11,1\%$	$30,8\%$	/	/
RLM	/	Test6CE	$0,58$	$\mathbf{35}\%$ $(6,2\%)$	$65,2\%$ $(6,3)$	$1,69$	$1,1$
		Test9CE	$0,65$	26% $(4,1\%)$	$52,7\%$ $(3,8\%)$	$2,36$	$1,32$
RLO	/	Test6CE	$\mathbf{0,59}$	$32,5\%$ $(5,3\%)$	$65,9\%$ $(7,7\%)$	$1,61$	$1,08$
		Test9CE	$0,67$	$26,3\%$ $(4,1\%)$	$53,5\%$ $(5,2\%)$	$2,2$	$\mathbf{1,28}$
Arbres de déc.	$cp = 0,003; minsplit = 10$	Test6CE	$0,53$	$30,3\%$ $(8,2\%)$	$60,4\%$ $(7,8\%)$	$1,88$	$1,17$
	$cp = 0,01; minsplit = 5$	Test9CE	$0,62$	22% $(1,8\%)$	$47,8\%$ $(4,8\%)$	$2,66$	$1,41$
Bagging	$nbagg = 150$	Test6CE	$0,51$	$28,8\%$ $(6,9\%)$	$64,5\%$ $(3,8\%)$	$1,72$	$1,13$
		Test9CE	$0,53$	21% $(3,9\%)$	$52,3\%$ $(6,6\%)$	$2,49$	$1,36$
Boosting	$iter = 40$	Test6CE	$0,58$	$32,9\%$ $(9,1\%)$	$\mathbf{68,5}\%$ $(9,5\%)$	$\mathbf{1,59}$	$\mathbf{1,07}$
		Test9CE	$0,63$	$24,3\%$ $(4,7\%)$	$\mathbf{55,5}\%$ $(6,2\%)$	$2,24$	$\mathbf{1,28}$
SVM	$\gamma = 0,05; C = 25$	Test6CE	$0,62$	$\mathbf{34}\%$ $(7,2\%)$	$\mathbf{68,2}\%$ (5%)	$\mathbf{1,51}$	$\mathbf{1,06}$
	$\gamma = 0,01; C = 40$	Test9CE	$\mathbf{0,68}$	$26,5\%$ $(4,2\%)$	$54,5\%$ $(4,8\%)$	$2,27$	$1,29$

Tableau 5.29 – Performances des différents modèles « baseline » sur Test6CE et Test9CE. Les méta-paramètres utilisés dans leur développement sont également repris.

Lorsqu'on compare ces différences entre modèles à l'aide du test-T pairé, il apparaît, sur *Test6CE,* que si les SVM sont significativement supérieures à RLM ($t(9) = 1,93; p = 0,042$) en termes d'$acc-cont$, ce n'est pas le cas au niveau de l'exactitude ($t(9) = 0.369; p = 0.64$). Par ailleurs, elles ne sont pas différentes de RLO, ni au niveau de l'exactitude ($t(9) = 0,97; p = 0,18$), ni d'$acc - cont$ ($t(9) = 1,11; p = 0,15$). On ne peut donc conclure strictement à la supériorité des SVM, même si elles se distinguent par un bon compromis entre

l'acc et l'$acc - cont$. C'est donc ce dernier modèle que nous avons retenu comme le meilleur en vue des comparaisons postérieures.

À noter que cette « baseline » obtient des scores à peine inférieurs à notre premier modèle (François, 2009b), lequel atteignait une exactitude de 41% et une $acc - cont$ de 71% sur un échantillon issu du *Corpus6*. Cela confirme donc bien la difficulté qu'il y a à dépasser la baseline classique, ici adaptée avec des algorithmes modernes, qui a été notée dans le Chapitre 2. Il est toutefois probable que ces performances relativement élevées de la « baseline » s'expliquent aussi par les nettoyages opérés sur nos corpus d'entraînement depuis 2009 (*cf.* Chapitre 4).

Le modèle Expert1

Ayant défini une solide « baseline », il convient désormais de chercher à la dépasser en recourant aux prédicteurs analysés dans la première partie de ce chapitre. La première tentative en ce sens a été effectuée en utilisant le meilleur prédicteur de chacune des quatre grandes familles, à savoir **PA_Alterego1a**, **NMP**, **avLocalLsa_Lem** et **BINGUI**. À nouveau, il a été nécessaire de rechercher les valeurs optimales pour les différents méta-paramètres des modèles. La procédure étant sensiblement la même que pour la « baseline », nous ne l'avons pas développée à nouveau. Les paramètres qui en résultent sont présentés au Tableau 5.30, tandis que les détails des expérimentations sont disponibles en annexes [26].

Modèle	Paramètres	Corpus	R	acc	$acc - cont$	$rmse$	mae
Baseline	/	Test6CE	$0,62$	34% $(7,2\%)$	$68,2\%$ (5%)	$1,51$	$1,06$
	/	Test9CE	$0,68$	$26,5\%$ $(4,2\%)$	$54,5\%$ $(4,8\%)$	$2,27$	$1,29$
RLM	/	Test6CE	$0,70$	$\mathbf{39}\%$ $(5,4\%)$	$74,2\%$ $(6,8\%)$	$1,34$	$0,97$
		Test9CE	$0,74$	$27,5\%$ $(7,1\%)$	$58,1\%$ $(5,9\%)$	$1,95$	$1,20$
RLO	/	Test6CE	$0,72$	$\mathbf{36},\mathbf{8}\%$ $(5,4\%)$	$\mathbf{77},\mathbf{8}\%$ $(5,3\%)$	$1,25$	$0,95$
		Test9CE	$\mathbf{0},\mathbf{77}$	$28,3\%$ $(4,6\%)$	$\mathbf{62},\mathbf{9}\%$ $(4,7\%)$	$1,79$	$\mathbf{1},\mathbf{15}$
Arbres de déc.	$minsplit = 10; cp = 0,01$	Test6CE	$0,65$	$33,4\%$ $(2,9\%)$	$69,3\%$ $(6,1\%)$	$1,6$	$1,07$
	$minsplit = 5; cp = 0,001$	Test9CE	$0,67$	$22,8\%$ $(3,2\%)$	$51,5\%$ $(4,2\%)$	$2,28$	$1,32$
Bagging	$nbagg = 200$	Test6CE	$0,66$	$32,7\%$ $(6,6\%)$	$72,9\%$ $(5,9\%)$	$1,40$	$1,01$
	$nbagg = 1000$	Test9CE	$0,69$	23% $(5,3\%)$	$55,2\%$ $(7,5\%)$	$2,09$	$1,26$
Boosting	$iter = 100, four$	Test6CE	$0,66$	34% $(7,6\%)$	74% $(6,2\%)$	$1,43$	$1,01$
	$iter = 75, four$	Test9CE	$0,69$	$23,8\%$ $(6,2\%)$	$56,4\%$ $(7,7\%)$	$2,06$	$1,25$
SVM	$\gamma = 0,006; C = 1$	Test6CE	$0,695$	35% $(7,6\%)$	$72,1\%$ $(5,4\%)$	$1,41$	$1,01$
	$\gamma = 0,1; C = 12$	Test9CE	$0,74$	$26,8\%$ $(7,2\%)$	58% $(5,7\%)$	$1,93$	$1,20$

Tableau 5.30 – Performances des différents modèles Expert1 sur Test6CE et Test9CE. Les méta-paramètres utilisés dans leur développement sont également repris.

Les résultats du Tableau 5.30 révèlent plusieurs faits intéressants. Tout d'abord, ils consacrent l'apparente supériorité des modèles logistiques sur le reste des algorithmes. Sur *Test6CE*,

26. Il convient de rapporter que, pour les SVM appliquées à **Dev6CE**, deux coûts très différents (1 et 1 000) ont été retenus, même après une seconde passe où γ a été maintenu constant. Nous avons donc décidé de tester ces deux ensembles de paramètres sur *Test6CE*. Leurs exactitudes étant très proches l'une de l'autre, c'est finalement un coût de 1 qui a été préféré.

RLM est significativement plus exacte que les SVM ($t(9) = 2,34; p = 0,02$), tandis que RLO dépasse ce même algorithme au niveau de l'$acc - cont$ ($t(9) = 3,18; p = 0,006$). De même, sur *Test9CE*, RLO classe en faisant moins d'erreurs graves [27] que les SVM de façon quasi significative ($t(9) = 1,61; p = 0,07$). Notons que l'écart-type des RLO est également plus réduit que celui des SVM. Par contre, on n'observe pas de différences significatives entre l'exactitude de RLM et celle des SVM ($t(9) = 0,44; p = 0,34$). Le fait que les SVM se comportent moins bien que pour la baseline pourrait s'expliquer par un mauvais choix des paramètres γ et C, car ayant appliqué par curiosité les valeurs obtenues sur le *Dev9CE* au modèle entraîné sur *Test6CE*, nous avons obtenu des résultats supérieurs à ceux présentés dans le Tableau 5.30. Quant aux autres modèles, ils obtiennent des résultats encore inférieurs, comme cela était déjà le cas du bagging et des arbres de décision sur la baseline. Notons que cette hiérarchie est parfaitement cohérente avec celle esquissée dans notre étude précédente (François, 2009a), alors que les variables utilisées sont pourtant différentes de celles du modèle de 2009.

C'est donc l'un des deux modèles logistiques qu'il faudrait retenir pour une formule courte à quatre facteurs pour le FLE, selon que ce soit l'exactitude ou l'exactitude contiguë qui prime. En effet, l'ajout de deux variables supplémentaires – respectivement relatives à la sémantique et au contexte de lecture en L2 – conduit bien à une amélioration générale des performances par rapport à la « baseline ». Ce gain de performance atteint presque les 10% en exactitude contiguë pour RLO ($t(9) = 4,14; p = 0,001$), avec pas moins de 78% au total, tandis que RLM gagne 5% d'exactitude sur le *Test6CE* ($t(9) = 1,77; p = 0,055$). Par ailleurs, soulignons que le modèle RLM, basé uniquement sur les quatre meilleurs prédicteurs, dépasse déjà les performances du modèle RLM de 2009 – qui comprenait plus de dix variables – en $acc - cont$ (74% contre 71%). L'exactitude de ce dernier reste cependant légèrement supérieure (41% contre 39%).

Ces résultats apparaissent donc en faveur d'une hypothèse essentielle de cette thèse, à savoir que la prise en compte de dimensions textuelles autres que lexico-syntaxiques permet d'améliorer les approches classiques. Toutefois, les scores obtenus en termes d'exactitude restent loin de ce que l'on pourrait espérer au sein d'une application réelle. C'est pourquoi nous avons cherché à améliorer *Expert1* en lui ajoutant quatre nouveaux facteurs, chacun issu de l'une des quatre dimensions textuelles.

27. Par le terme d'« erreur grave », nous désignons une prédiction qui s'éloigne au moins de deux niveaux de l'annotation manuelle attribuée au texte, tel que classer un document A2 alors qu'il est B2.

Le modèle Expert2

Le second et dernier modèle expert envisagé dans cette étude rassemble quant à lui les deux meilleurs prédicteurs de chaque famille : **PA_Alterego1a**, **X90FFFC**, **NMP**, **PPres**, **avLocalLsa_Lem**, **PP1P2**, **BINGUI** et **NAColl**, ce qui en fait huit. Il vise à vérifier si l'ajout d'un plus grand nombre de variables conduit à une amélioration des capacités de discrimination, comme le pensait Bormuth (1966), ou si la parcimonie classique devrait également être respectée dans le contexte de notre étude. Dans ce cas, le meilleur modèle Expert1 devrait rester supérieur au meilleur modèle Expert2.

Comme précédemment, il a été nécessaire de déterminer une série de méta-paramètres sur les corpus de développement *Dev6CE* et *Dev9CE*. Les valeurs obtenues au terme du processus de « grid search » sont rapportées au Tableau 5.31, de même que les résultats des différents algorithmes sur les corpus de test. À nouveau, ce sont les trois mêmes approches qui ressortent : les SVM, la RLO et la RLM. Cette fois, ce sont les SVM qui obtiennent les meilleurs scores sur le *Test6CE* et la meilleure exactitude sur le *Test9CE*. Cependant, il n'y a pas de différence significative entre la RLM et les SVM, ni au niveau de l'exactitude, ni au niveau de l'$acc - cont$ ($t(9) = 0,43; p = 0,34$ pour *Test6CE* et $t(9) = 0,29; p = 0,39$ pour *Test9CE*). Quant à la meilleure $acc - cont$ sur *Test9CE*, elle est obtenue par la RLO, sans que cette supériorité ne soit significative par rapport aux SVM ($t(9) = 1,53; p = 0,08$). Elle l'est toutefois par rapport à la RLM ($t(9) = 2,89; p = 0,009$).

Modèle	Paramètres	Corpus	R	acc	$acc - cont$	$rmse$	mae
Baseline	/	Test6CE	0,62	34% (7,2%)	68,2% (5%)	1,51	1,06
	/	Test9CE	0,68	26,5% (4,2%)	54,5% (4,8%)	2,27	1,29
RLM	/	Test6CE	0,73	**41**% (4,4%)	**77**% (5,1%)	1,31	0,95
		Test9CE	0,75	31% (6,5%)	61,7% (4,5%)	1,92	1,17
RLO	/	Test6CE	0,74	**40**% (11,1%)	76,6% (4,2%)	1,25	0,94
		Test9CE	**0,77**	29,5% (3,7%)	**66,5**% (5,5%)	1,77	**1,13**
Arbres de déc.	$minsplit = 10; cp = 0,007$	Test6CE	0,63	35,1% (6,7%)	68,8% (7,7%)	1,59	1,06
	$minsplit = 20; cp = 0,0005$	Test9CE	0,68	25,4% (3,9%)	51,2% (6,3%)	2,23	1,29
Bagging	$nbagg = 175$	Test6CE	0,70	37,8% (7%)	75,6% (5,4%)	1,33	0,97
	$nbagg = 250$	Test9CE	0,72	27,7% (4%)	60,5% (5,7%)	1,95	1,20
Boosting	$iter = 40, four$	Test6CE	0,70	38% (7,4%)	76% (5,4%)	1,33	0,97
	$iter = 75, four$	Test9CE	0,73	30,3% (6,2%)	**63,5**% (5,6%)	1,94	1,18
SVM	$\gamma = 0,002; C = 75$	Test6CE	0,73	**41**% (5,2%)	**78**% (4,1%)	1,28	0,94
	$\gamma = 0,006; C = 20$	Test9CE	0,75	**31**% (5,5%)	62,3% (6,3%)	1,90	1,17

Tableau 5.31 – Performances des différents modèles Expert2 sur Test6CE et Test9CE. Les méta-paramètres utilisés dans leur développement sont également repris.

En résumé, la RLO pourrait être utilisée si l'exactitude contiguë était le seul critère, mais mieux vaut retenir les SVM, qui combinent les scores les plus élevés en exactitude et en exactitude contiguë. Sur la base de ce modèle, on note que l'ensemble de variables *Expert2* obtient des meilleures performances que *Expert1*. Si cette différence est insuffisante pour conclure à la supériorité d'*Expert2* sur le *Test6CE*, elle est par contre tout à fait significative

sur le *Test9CE* [28]. Par conséquent, il n'est pas étonnant qu'*Expert2* produise des modèles qui sont également largement plus performants que la « baseline » [29].

Dès lors, il semblerait bien que l'extrême parcimonie classique ne soit plus un exemple à suivre aujourd'hui, puisque notre formule à huit prédicteurs gagne tout de même entre 5% à 7% en exactitude et entre 8% et 10% en exactitude contiguë par rapport à la « baseline ».

Par contre, il apparaît que notre remarque précédente, arguant la nécessité de marier des indices de lisibilités issus de dimensions textuelles les plus diversifiées possible, doit être remise en question. En effet, le modèle *Expert2*, basé sur une analyse systématique d'un grand nombre de prédicteurs, ne conduit pas à des prédictions plus exactes que notre premier modèle (François, 2009b), lequel n'intégrait aucune variable sémantique. Il semble donc que le nombre de variables puisse jouer un rôle aussi crucial que leur qualité, pour autant que la quantité d'information différente apportée au modèle soit optimisée. On notera à ce propos que le modèle de 2009 a été obtenu suite à une sélection automatique de variables parmi 20 candidats. Toutefois, cette problématique ne peut être clairement tranchée, vu que l'on observe tout de même un gain de 8% d'$acc - cont$ pour *Expert2* par rapport au modèle de 2009. Il est donc possible que ce soit surtout cette seconde dimension qui profite de la diversification des dimensions textuelles considérées.

Ce constat laisse néanmoins penser que la sélection des variables par expertise n'est peut-être pas la meilleure des options. Tout d'abord, le fait de se restreindre aux deux meilleures variables par famille a conduit à sélectionner certains prédicteurs (en particulier **NAColl**) dont l'efficacité, bien qu'avérée au sein de leur famille, reste faible en regard de celle de prédicteurs d'autres familles, écartés et qui auraient pu se révéler plus informatifs. D'autre part, ce choix revient à optimiser la quantité totale d'information, sans s'assurer que celle-ci n'est pas redondante. Le Tableau 5.32, qui montre les corrélations croisées entre les huit indices d'*Expert2*, révèle d'ailleurs plusieurs variables relativement redondantes. Parmi celles-ci, il est étonnant de noter que la LSA semble très proche de **NMP**, une variable syntaxique. Cela expliquerait toutefois pourquoi l'ajout de cet indice si prometteur n'a pas permis d'élever autant les performances de notre formule que nous ne l'espérions.

Afin de déterminer plus précisément l'apport de chaque variable une fois que les autres

28. Pour effectuer les tests de significativité, nous avons retenu la RLM comme représentant d'*Expert1*. Sur le *Test6CE*, la différence d'exactitude est non significative ($t(9) = 0,66; p = 0,26$), alors que la différence d'exactitude contiguë est à la limite de la significativité ($t(9) = 1,76; p = 0.056$). Par contre, sur *Test9CE*, la différence entre les deux modèles est significative, que ce soit au niveau d'acc ($t(9) = 2,17; p = 0,03$) comme pour $acc - cont$ ($t(9) = 2,1; p = 0,03$).

29. Sur le *Test6CE*, la différence d'exactitude entre la baseline et les SVM entraînées sur *Expert2* est de 7%, ce qui est tout à fait significatif ($t(9) = 2,36; p = 0,02$). Il en va de même pour les 10% de gain en exactitude contiguë ($t(9) = 4,05; p = 0,0014$). Quant à *Test9CE*, la supériorité du modèle à huit variables ne laisse aucun doute, ni pour acc ($t(9) = 2,08; p = 0,034$), ni pour $acc - cont$ ($t(9) = 3,25; p = 0,005$).

	Diff	NMP	X90FFFC	avLoc.	PA_Alt.	PPres	PP1P2	BINGUI	NAColl
Diff	1	$0,57$	$-0,61$	0.55	$0,60$	$0,47$	$-0,30$	$0,45$	$0,21$
NMP	/	1	$\mathbf{-0,51}$	$\mathbf{0,60}$	$0,47$	$0,36$	$-0,42$	$0,39$	$0,18$
X90FFFC	/	/	1	$-0,49$	$\mathbf{-0,68}$	$0,44$	$\mathbf{0,47}$	$-0,35$	$-0,19$
avLocalLsa_Lem	/	/	/	1	$0,46$	$0,34$	$-0,39$	$0,33$	$0,14$
PA_Alterego1a	/	/	/	/	1	$0,35$	$-0,40$	$0,27$	$0,23$
PPres	/	/	/	/	/	1	$-0,14$	$0,27$	0.23
PP1P2	/	/	/	/	/	/	1	$-0,17$	$-0,08$
BINGUI	/	/	/	/	/	/	/	1	0.11
NAColl	/	/	/	/	/	/	/	/	1

Tableau 5.32 – Corrélations croissées entre les variables de l'ensemble Expert2, calculées sur Test6CE.

sont déjà présentes, nous avons entraîné 8 modèles par SVM comprenant $p - 1$ variables (en gardant les méta-paramètres constants) et les avons comparé au modèle complet. Cela donne le gain en acc et en $acc - cont$ pour chaque variable une fois que les autres sont déjà présentes, ce qui revient à peu près à calculer une corrélation partielle (Howell, 2008, 527). L'ensemble de ces gains sont repris dans les Figures 5.12 pour le modèle à 6 classes et 5.13 pour celui à 9 niveaux.

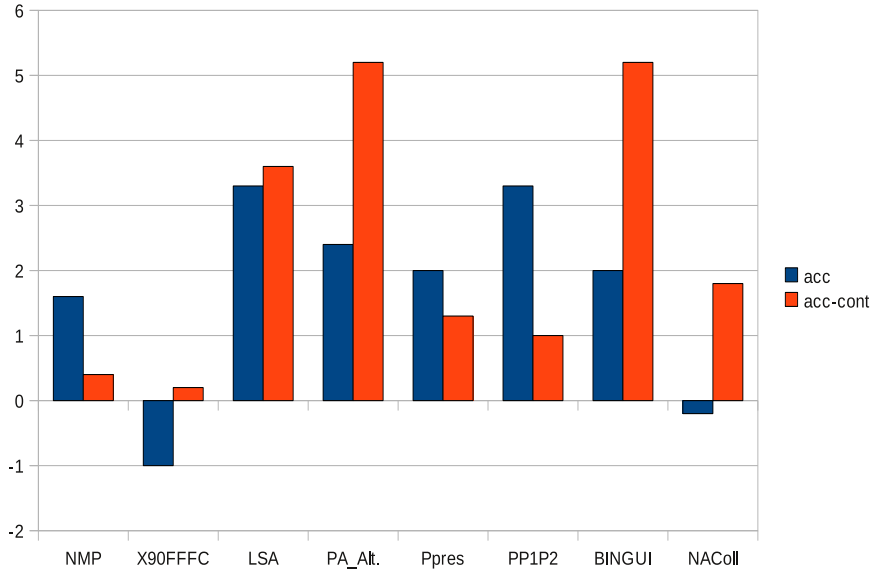


Figure 5.12 – Différences entre le modèle Expert2 complet et les modèles excluant l'une des variables, calculées par classifieur SVM sur Test6CE. Plus la différence est positive, plus la variable associée apporte une information nouvelle au modèle.

On y fait plusieurs constats intéressants. Tout d'abord, il est clair que certains prédicteurs apportent plus d'information au modèle que d'autres. C'est le cas, en particulier, de **PA_Alterego1a**. Non content d'être le meilleur prédicteur envisagé seul, il est également celui qui améliore le plus l'exactitude contiguë des modèles (à 6 comme à 9 classes) lorsqu'il est le dernier à

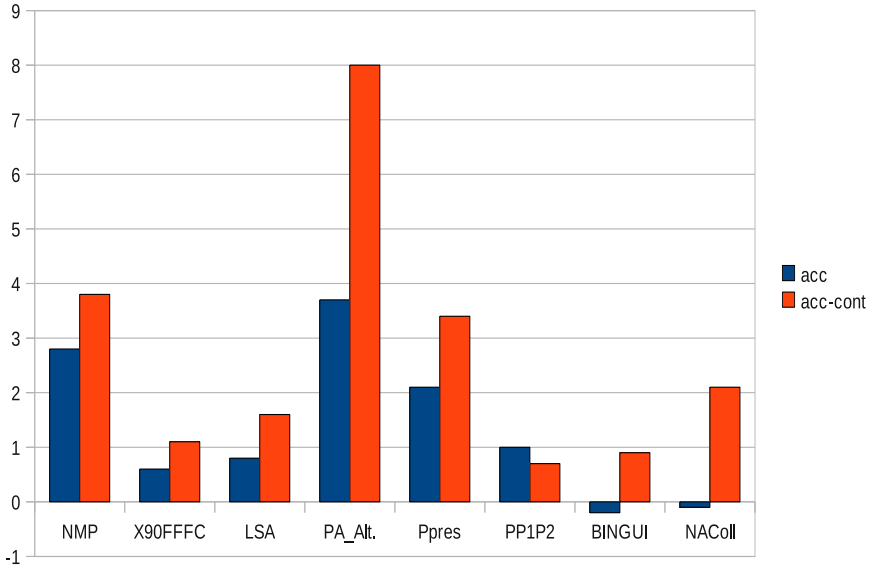


Figure 5.13 – Différences entre le modèle Expert2 complet et les modèles excluant l'une des variables, calculées par classifieur SVM sur Test9CE. Plus la différence est positive, plus la variable associée apporte une information nouvelle au modèle.

être inclus. Ces résultats confirment une fois de plus l'excellence de cette variable, quand bien même **X90FFFC**, avec lequel il est corrélé à $\rho = -0,68$, est déjà présent.

Ensuite, on remarque que certains indices favorisent presque exclusivement l'exactitude (**PP1P2**) ou l'$acc - cont$ (**BINGUI** et **NAColl**), tandis que d'autres ont un effet conjoint sur ces deux dimensions (**avLocalLsa_Lem** ou **NMP**). Par ailleurs, l'effet de certaines variables varie sensiblement entre les deux modèles, possiblement en raison de différences au niveau des textes constitutifs des corpus. Enfin, certains aspects textuels apparaissent peu utiles pour la formule, en particulier **X90FFFC**, dont la majorité de l'information qu'il apporte semble déjà couverte par d'autres prédicteurs. Cette analyse révèle donc l'importance de prendre en compte les corrélations croisées entre variables, puisqu'à la lecture de ces résultats, on pourrait reformuler un modèle qui se passerait de **X90FFFC**, voire aussi de **BINGUI** pour *Test9CE*.

Conclusion

Cette section a présenté divers modèles à base d'une sélection manuelle des prédicteurs, laquelle repose sur le postulat théorique qu'une bonne formule de lisibilité devrait comprendre des prédicteurs correspondant à divers niveaux textuels. Pour vérifier la validité de cette hypothèse, une « baseline » d'inspiration classique, utilisant les SVM, a tout d'abord été sélectionnée comme point de référence pour les modèles conçus à partir de deux ensembles de variables plus larges. Suite à la comparaison entre ces trois ensembles de prédic-

teurs, il est apparu qu'augmenter le nombre de variables prises en compte par les modèles, du moins jusqu'à huit, conduit bien à des améliorations de performances significatives. Ces résultats abondent donc dans le sens de Bormuth (1966) concernant l'intérêt de formules plus complexes que celles recommandées par l'idéal classique.

Par contre, il n'a pas été possible de déterminer avec certitude si ce gain de performance par rapport à la « baseline » provient de la diversification des dimensions textuelles considérées ou simplement d'une augmentation du nombre de prédicteurs. En effet, l'application des SVM à *Expert2* a conduit à des performances similaires, en termes d'exactitude, à celles de notre formule de 2009, qui ne comprenait pourtant pas de variables sémantiques. Certes, une amélioration a pu être observée pour $acc - cont$, mais celle-ci avait déjà été obtenue pour le modèle *Expert1* par RLO. Une observation supplémentaire est donc que l'ajout de nouveaux prédicteurs semble davantage influencer l'exactitude du modèle que son exactitude contiguë.

Afin de chercher à trancher cette question de l'apport des différentes dimensions, mais également afin de chercher à dépasser le meilleur modèle *Expert2*, une approche de sélection de variables, basée sur la maximisation de l'information nouvelle apportée, est appliquée en amont de la modélisation dans la section qui suit.

5.4.5 Comparaison des modèles par sélection automatisée de variables

Introduction

Dans cette seconde partie de l'étape de modélisation, une approche alternative de la sélection des variables est adoptée : automatiser cette procédure à l'aide des techniques décrites à la Section 5.3.3. Comme cette sélection automatique dépend fortement des caractéristiques de chaque modèle, elle est discutée séparément pour chacun des six algorithmes statistiques. En effet, certains, telle la régression logistique, nécessitent la mise en place d'une procédure préalable de sélection ; d'autres, comme les arbres de décision et ses dérivés, intègrent cette étape dans leur conception, tandis que d'autres encore (les SVM) sont parfois employés sans aucune sélection.

Toutefois, nous avons été confronté à un problème touchant plusieurs modèles. Au sein de chaque famille, les variables sont généralement fort similaires les unes les autres, étant donné qu'elles représentent des variantes d'un nombre limité d'hypothèses. Dès lors, la matrice des données qui est produite à la fin de l'étape de paramétrisation souffre d'un problème de colinéarité, c'est-à-dire que certains prédicteurs peuvent être obtenus à partir de la combinaison linéaire d'une ou plusieurs autres variables. Or, l'entraînement des paramètres des modèles logistiques est touché par cette caractéristique, vu qu'il requiert d'inverser $\mathbf{R}$, la matrice de

corrélation entre les p prédicteurs, et qu'une matrice souffrant de colinéarité ne peut être inversée (Foucart, 2006, 6).

Dès lors, il a fallu envisager plusieurs solutions pour contourner ce problème. La plus importante d'entre elles a consisté à opérer un tri parmi les 406 variables, comme le conseillent Hosmer et Lemeshow (1989) (*cf.* Section 5.3.3), à l'aide des analyses rapportées dans la première partie de ce chapitre. Ainsi, seules les 47 variables suivantes ont finalement été sélectionnées au sein de chaque sous-famille :

- **Fréquences lexicales** : MeanFFFC, MedianFFFC, IQFFFC, X75FFFC et X90FFFC ;
- **Pourcentage d'absents** : PAGoug_2000 et PA_Alterego1a ;
- **Modèle unigramme** : ML3 ;
- **Modèles n-grammes** : meanNGProb.G ;
- **Indices de la diversité lexicale** : TTR_W ;
- **Le voisinage orthographique** : freqCumNeigh_75, nbNeighMoreFreq_90, med_nbnbNeighMoreFreq ;
- **Indices de la longueur des mots** : NLM, PM8, NL90P ;
- **Mesures de la longueur syntaxique** : NMP, NWS_902 , LenSentDaoust2 ;
- **Ratio des catégories grammaticales** : PRE.ALL, GRAM.PRO, PRE.PRO, ART.PRO, PRE.LEX, PRE.GRAM, ART.LEX, NOM.NAM.ART ;
- **Temps verbaux** : tous les indices binaires (sauf Pres, qui est remplacé par Pres_C) ;
- **Le taux personnalisation** : PP1P2, PP2 ;
- **La densité des idées** : ConcDens ;
- **Le taux de cohérence** : avLocalLsa_Lem ;
- **Les variables basées sur les unités polylexicales** : NCPW, NAColl, PColl90 ;
- **Indice du type texte** : PPD, BINGUI.

La régression logistique ordinale

Pour la RLO, nous avons employé un algorithme de sélection « pas à pas » de type descendant, fourni par le module *Design* de R (Alzola et Harrell, 2002) et reposant sur le critère AIC (*cf.* Section 5.3.3). Il présente par ailleurs l'avantage de réitérer M fois la sélection sur des corpus légèrement différents obtenus via une procédure de « bootstrapping .632 » (Tufféry, 2007, 369-371). Cette répétition permet d'éliminer la plupart des effets dus au hasard, à cause desquels une variable peu utile dans la plupart des cas peut néanmoins être retenue pour un corpus particulier. Il est apparu que 100 répétitions suffisaient déjà largement pour obtenir cette stabilité dans la sélection.

Par contre, malgré la sélection drastique opérée parmi les 406 variables, le problème de colinéarité n'a pas été complètement réglé pour autant. Il a donc fallu recourir à une seconde

stratégie qui consiste à effectuer une régularisation des paramètres au cours de l'entraînement du modèle (Bishop, 2006, 10). Celle-ci a comme effet de réduire la taille des coefficients du modèle et diminue aussi les problèmes de colinéarité au sein des données. La pénalité que nous avons employée est de $0,1$, l'objectif étant de la conserver aussi petite que possible.

Les résultats obtenus à l'aide de cette procédure sont assez surprenants sur *Dev6CE,* car seules deux variables sont conservées : **PA_Alterego1a** et **NMP**. Il s'agit certes de la meilleure des prédicteurs lexicaux et de l'une des plus efficaces au niveau syntaxique, mais on se serait attendu à un choix moins sélectif. Toutefois, notons que **PPasse** a tout de même été retenu 41 fois sur les 100 itérations et que **ML3** l'a été à 50 reprises, alors que les autres facteurs n'ont été conservés qu'épisodiquement. C'est pourquoi nous avons décidé d'évaluer deux modèles différents pour la RLO à 6 classes, car celui obtenu automatiquement risque fort de ne pas être significativement plus efficace que la « baseline » :

– **AutoOLR6-1** : PA_Alterego1a + NMP
– **AutoOLR6-2** : PA_Alterego1a + NMP + PPres + ML3

Soulignons que le premier ressemble fort à la formule de Dale et Chall (1948a), à ceci près que la liste de 3000 mots de Dale y est remplacée par un lexique propre au FLE. Il est également fort proche de notre « baseline », **PA_Alterego1a** s'étant simplement substitué à **NLM**. Ce résultat confirme la solidité du duo lexico-syntaxique pour les formules de lisibilité. Quant au second modèle, il y ajoute deux variables dont nos travaux précédents François (2009a) ont mis en évidence l'efficacité.

Pour la formule à 9 niveaux, la procédure de sélection automatique a estimé nécessaire d'accepter davantage de variables afin de traiter un problème rendu plus ardu par la présence d'un plus grand nombre de classes. Le modèle retenu est le suivant :

AutoOLR9 : ML3 + NMP + Impf + PA_Alterego1a + nbNeighMoreFreq_90 + meanNGProb.G

On y retrouve la même base que pour la formule à 6 classes, ainsi que **ML3** et une variable de temps (**Impf**). Il est par contre plus étonnant de retrouver des prédicteurs reposant sur les modèles n-grammes et même les voisins orthographiques, au vu de leur piètre capacité de discrimination mesurée via l'analyse bivariée.

La régression logistique multinomiale

Pour la RLM, le module *nnet* du logiciel R que nous avons utilisé n'offrant pas de technique de sélection automatique de variables, c'est celle du module *MASS* (Venables et Ripley, 2002) qui a été employée. Elle repose également sur une procédure « pas-à-pas » descendante guidée par AIC. Toutefois, elle apparaît moins sélective que la précédente, ce qui est d'autant plus étonnant que la RLM comportant plus de paramètres que la RLO, la suppression d'une variable diminue pourtant davantage la valeur du critère AIC. Les deux modèles retenus sont les suivants :

AutoMLR6 : PA_Alterego1a + Cond + Imperatif + Impf + PPasse + PPres + Subi + Subp + BINGUI + TTR_W + NWS_902 + LenSentDaoust2 + med_nbNeighMoreFreq

AutoMLR9 : PA_Alterego1a + PPD + BINGUI + Impf + PPasse + PPres + TTR_W + NWS_902 + nbNeighMoreFreq_90 + NOM.NAM.ART

On retrouve à nouveau l'excellente variable **PA_Alterego1a** dans les deux sélections, lesquelles comprennent aussi **NWS_902** qui s'était révélé à peu près aussi efficace que **NMP** dans l'analyse bivariée (*cf.* Section 5.2.5). Comme pour le modèle de François (2009a), plusieurs variables temporelles sont retenues, confirmant l'intérêt de ce type d'information. Par contre, l'entrée de **NOM.NAM.ART**, qui marque la première apparition d'un ratio de catégorie grammaticale, est plus inattendue, de même que celle de **TTR_W**. Remarquons aussi que, à l'inverse de la RLO, la RLM à 6 niveaux comprend plus de prédicteurs (à savoir 13) que la formule à 9 classes, laquelle n'en intègre que 10.

Les arbres de décision et les modèles dérivés

Comme il l'a été précisé à la Section 5.3.3, les arbres de décision ne nécessitent pas le recours à une procédure spéciale de sélection de variables, étant donné qu'ils opèrent celle-ci à chaque fois qu'un nœud est divisé. Dès lors, l'entraînement du modèle est similaire à celui de la section sur la sélection manuelle des prédicteurs, à ceci près que l'algorithme a désormais la possibilité d'aller piocher dans les 47 variables précitées plutôt que les quatre ou huit des approches *Expert1* et *Expert2*. Comme précédemment, les échantillons de développement ont servi à calculer des valeurs optimales pour $minsplit$ et cp, qui sont rapportées au Tableau 5.33. Il en va de même pour le bagging et le boosting, qui utilisent des arbres de décision comme classifieurs, et dont les méta-paramètres retenus sont aussi précisés dans le Tableau 5.33.

Les séparateurs à vaste marge

Enfin, pour les SVM, nous avons employé l'ensemble des variables sans opérer de sélection[30]. Bien qu'il existe diverses techniques permettant d'effectuer celle-ci (Chen et Lin, 2006; Guyon *et al.*, 2002), les SVM utilisant déjà une régularisation de base, elles sont moins sensibles aux problèmes de dimensionalité (Guyon *et al.*, 2002, 391). Ces derniers auteurs ont certes montré que, lorsque le nombre de prédicteurs est très grand (plusieurs milliers) et le nombre d'observations limité à quelques dizaines, les SVM profitent d'une réduction de dimensionalité. Cette situation n'étant pas la nôtre, il n'a pas semblé nécessaire de recourir à ce type de techniques, d'autant que l'adaptation multi-classes de l'algorithme rend leur application nettement plus complexe (Chapelle et Keerthi, 2008). Cette façon de faire est de toute façon assez classique dans la littérature et n'a pas engendré de perte de performances (au contraire) pour les 46 variables utilisées. Comme précédemment, les méta-paramètres rapportés dans le Tableau 5.33 ont été obtenus via une « grid search » basée sur une validation croisée à 10 échantillons.

Discussion des résultats

Le comportement de ces différents modèles est repris dans le Tableau 5.33 et se révèle extrêmement intéressant à plusieurs égards. Tout d'abord, on remarque que les algorithmes qui sortent du lot ne sont plus exactement les mêmes que pour l'approche experte. En effet, la RLM et la RLO se voient surpassées par le boosting et le bagging, tandis que les SVM demeurent parmi les meilleurs classifieurs.

Plus précisément, le boosting apparaît comme le meilleur modèle pour le *Test9CE*, avec rien moins que 37% d'exactitude et $69,4\%$ d'exactitude contiguë. Il s'avère significativement meilleur que la RLM, aussi bien en termes d'exactitude que d'$acc - cont$[31]. De même, le gain de 7% par rapport à la RLO à 9 niveaux est extrêmement significatif ($t(9) = 5,5; p = 0.0002$), même si la différence d'$acc - cont$ n'est pas suffisante quant à elle ($t(9) = 1,27; p = 0,12$). Ces tests confirment donc la supériorité du boosting sur la régression logistique, à la différence des modèles experts.

En ce qui concerne les modèles à 6 niveaux, ce sont cette fois les SVM qui dominent. Si les $3,5\%$ d'exactitude contiguë par rapport à la RLO et les $3,9\%$ par rapport à la RLM

30. Soulignons que **PColl90** a néanmoins dû être retiré de cet ensemble, car le nombre important de valeurs manquantes pour cette variable (64 sur 408) nuisait à la stabilité du modèle. C'est pour cette raison que référence est faite plus loin à 46 et non 47 prédicteurs.

31. Pour l'exactitude, le niveau de significativité est de $t(9) = 2,54; p = 0,016$, tandis qu'il est de $t(9) = 2,15; p = 0,03$ pour $acc - cont$. Par contre, le gain du modèle par boosting à 6 niveaux n'est pas suffisant pour atteindre le seuil de significativté : acc : $t(9) = 0.5565; p = 0,30$ et $acc - cont$: $t(9) = 0,31; p = 0,38$.

ne font que frôler la significativité (respectivement $t(9) = 1,55; p = 0,08$ et $t(9) = 1,59; p = 0,07$), les différences d'exactitude sont par contre plus nettes (respectivement $t(9) = 2,31; p = 0,02$ et $t(9) = 1,79; p = 0,05$). Les performances du boosting sont quant à elles suffisamment proches de celles des SVM pour que les deux modèles soient statistiquement équivalents.

Modèle	Paramètres	Corpus	R	acc	$acc - cont$	$rmse$	mae
Baseline	/	Test6CE	$0,62$	34% $(7,2\%)$	$68,2\%$ (5%)	$1,51$	$1,06$
	/	Test9CE	$0,68$	$26,5\%$ $(4,2\%)$	$54,5\%$ $(4,8\%)$	$2,27$	$1,29$
RLM	/	Test6CE	$0,71$	44% $(5,8\%)$	$75,7\%$ $(5,3\%)$	$1,32$	$0,95$
		Test9CE	$0,74$	33% $(3,6\%)$	$64,5\%$ (4%)	$1,91$	$1,16$
RLO	/	Test6CE-1	$0,68$	34% $(5,5\%)$	$71,5\%$ $(4,9\%)$	$1,41$	$1,01$
		Test6CE-2	$0,71$	40% $(7,2\%)$	$76,1\%$ $(5,3\%)$	$1,33$	$0,96$
		Test9CE	$\mathbf{0,77}$	$30,4\%$ $(3,2\%)$	$66,2\%$ (6%)	$1,76$	$\mathbf{1,13}$
Arbres de déc.	$minsplit = 2; cp = 0.013$	Test6CE	$0,63$	37% $(7,4\%)$	$71,2\%$ (7%)	$1,54$	$1,04$
	$minsplit = 10; cp = 0,001$	Test9CE	$0,71$	$31,4\%$ $(5,4\%)$	$60,7\%$ $(5,1\%)$	$2,06$	$1,21$
Bagging	$nbagg = 100$	Test6CE	$0,71$	$44,2\%$ $(7,1\%)$	$75,6\%$ $(6,5\%)$	$1,34$	$0,94$
	$nbagg = 250$	Test9CE	$\mathbf{0,78}$	$\mathbf{36,3\%}$ $(4,7\%)$	$\mathbf{70\%}$ (7%)	$1,77$	$1,10$
Boosting	$iter = 30, four$	Test6CE	$0,71$	$\mathbf{46\%}$ (10%)	$\mathbf{76,6\%}$ $(8,4\%)$	$1,32$	$0,93$
	$iter = 150, four$	Test9CE	$0,76$	$\mathbf{37\%}$ (3%)	$\mathbf{69,4\%}$ $(6,4\%)$	$1,81$	$1,11$
SVM	$\gamma = 0,004; C = 5$	Test6CE	$0,73$	$\mathbf{49\%}$ $(9,7\%)$	$\mathbf{79,6\%}$ $(5,4\%)$	$1,27$	$0,90$
	$\gamma = 0,004; C = 15$	Test9CE	$0,74$	35% $(7,4\%)$	$65,4\%$ $(7,4\%)$	$1,92$	$1,15$

Tableau 5.33 – Performances des différents modèles automatiques sur Test6CE et Test9CE. Les méta-paramètres utilisés dans leur développement sont également repris.

On dénote donc un changement dans le comportement des modèles, par rapport à leur version *Expert*. Pour les seconds, qui comportent à la fois moins de variables et une information moins complémentaire, la régression logistique a tendance à dominer ou à être aussi efficace qu'un algorithme plus complexe tel que les SVM. Le choix de la régression logistique est dès lors celui de la parcimonie. Ce résultat réplique d'ailleurs ceux de plusieurs études précédentes (Feng *et al.*, 2010; François, 2009a,b; Heilman *et al.*, 2007).

Cependant, déjà chez Feng *et al.* (2010), on remarque que si la régression logistique et les SVM atteignent des scores relativement similaires lorsqu'ils sont entraînés avec une même famille de variables, les SVM obtiennent systématiquement des résultats légèrement supérieurs dès que des variables différentes sont employées en plus grand nombre. Or, c'est très exactement ce motif, plus accentué encore, que nous obtenons. Il apparaît nettement que, lorsque les modèles complexes, tels que le boosting, le bagging et les SVM, ont à leur disposition un plus large ensemble de prédicteurs, ils parviennent à en faire un meilleur usage, lequel se marque en particulier au niveau de l'exactitude.

En effet, on observe cette fois des différences très nettes entre la « baseline » et le meilleur modèle automatique pour chacune des deux échelles de difficulté. Sur *Test6CE*, les SVM amènent un gain de 15% en exactitude, tandis que le boosting permet de gagner $10,5\%$ sur *Test9CE*. Il s'agit également d'un mieux par rapport à nos modèles de 2009, qui atteignaient respectivement 41% et 32% d'exactitude, de même que par rapport au meilleur modèle *Ex-*

pert2 (avec 41% et 31%). Les modèles au sein desquels la sélection de variables est opérée automatiquement supplantent donc effectivement les modèles experts, probablement parce qu'ils permettent de mieux optimiser l'apport d'information des différents prédicteurs.

Conclusion

On peut donc tirer plusieurs conclusions de ces expériences. Tout d'abord, le fait de chercher à diversifier au maximum l'information apportée au modèle a effectivement conduit à des améliorations significatives des modèles, contrairement aux simples expérimentations de nouveaux algorithmes statistiques. Nous répondons ainsi à une question soulevée dans nos travaux précédents (François, 2009b), à savoir qu'il semble être plus efficace de concentrer les recherches en lisibilité sur la découverte de nouveaux prédicteurs que sur des optimisations statistiques.

Toutefois, la réponse à cette question n'est pas aussi simple. En effet, nos expériences ont révélé la possibilité d'une interaction entre le type de modèles employés et le nombre P de variables qu'ils intègrent. La régression logistique se comporte mieux lorsque P est réduit, mais est supplantée par des classifieurs plus complexes tandis que P augmente. Cette observation n'avait pas encore été rapportée dans la littérature, à notre connaissance, et mériterait d'être explorée plus en détail.

Par ailleurs, il apparaît que l'accroissement de P conduit principalement à une hausse de l'exactitude, mais a nettement moins d'effet sur l'exactitude contiguë, où le gain de performance est faible entre les modèles *Expert1*, *Expert2* et le modèle obtenu automatiquement. Cela signifie que ce dernier type de classifieurs est plus exact, mais commet toujours le même nombre de fautes graves.

Enfin, les scores très moyens du modèle RLO à deux variables confirment à notre sens les limites de l'approche classique. Quelles que soient les deux ou trois variables employées, celles-ci ne suffisent pas à dépasser un certain seuil d'efficacité, que les approches modernes combinant le TAL et les techniques d'apprentissage automatisé peuvent prétendre atteindre. Ces résultats nous semblent donc bien démontrer, une nouvelle fois, l'apport du TAL à la lisibilité du FLE, qui était l'une des questions principales de cette thèse. Ils débouchent également sur le choix d'une nouvelle formule de lisibilité pour le FLE, sur laquelle nous allons revenir dans le point suivant.

5.4.6 La formule finale

Introduction

Au terme de cette comparaison de divers modèles, tantôt définis par expertise, tantôt à l'aide de méthodes automatiques, deux modèles sont apparus comme les plus efficaces pour classer des textes de FLE en fonction de l'échelle de difficulté du CECR : le boosting et les SVM. Le premier d'entre eux se comporte mieux sur le corpus à 9 classes, tandis que le second s'est distingué sur le problème à 6 classes. Toutefois, ces légères différences d'efficacité ne se sont pas révélées significatives, ce qui rend indifférent l'emploi de l'un ou l'autre de ces deux algorithmes.

L'aspect déterminant dans le choix d'une formule finale a donc été le temps requis pour son entraînement. Certes, les deux approches requièrent l'apprentissage d'un grand nombre de modèles. Les SVM pratiquant le « un contre un », cela nécessite d'apprendre 15 ou 36 classifieurs binaires. Toutefois, cela n'est encore rien en comparaison du boosting qui, à chacune de ses nombreuses itérations, utilise six classifieurs binaires. Dès lors, il est clair que cette seconde technique demande davantage de temps. Sur nos données, on passe de quelques secondes pour les SVM à plus de deux heures d'entraînement. Par conséquent, il nous semble plus judicieux de retenir les SVM comme classifieurs pour notre formule finale.

Partant de cet acquis, cette section clôt notre étude en analysant plus précisément les performances du modèle retenu, proposant des diagnostics complémentaires aux cinq mesures d'évaluation rapportées dans le Tableau 5.33 et comparant l'efficacité de celui-ci par rapport à celle d'autres formules de lisibilité computationnelle.

La formule retenue et ses performances

Le résumé des deux formules retenues au terme de cette étude est présenté dans le Tableau 5.34. Malheureusement, il n'est guère possible de donner plus de détails sur l'agencement des variables ou leurs coefficients, car les SVM ne constituent pas un modèle aussi interprétable qu'une régression logistique. Cette limitation n'est toutefois pas cruciale étant donné que la visée de cette étude était de développer un modèle prédictif et non un modèle explicatif.

Ce qui importe dès lors, ce sont les capacités de prédiction de cette formule sur des nouveaux textes. Celles-ci se situent autour 49% pour l'échelle originale du CECR et autour de 35% pour l'échelle raffinée. Cela peut sembler peu pour une tâche dans laquelle, s'il est possible de pardonner une légère imprécision, les erreurs graves sont difficilement tolérables. En effet, quelle est l'utilité d'un système qui classe des textes B2 comme étant A2, sachant que

	Modèle à 6 classes	**Modèle à 9 classes**
Algorithme	SVM « un contre un »	SVM « un contre un »
SVM-Type	C-classification	C-classification
Cost	5	15
Gamma	0,004	0,004
Nb. de vecteurs supports	±335	±500
Nb. de variables	46	46
Nb. de données d'entraînement	398	592
R	0,73	0,74
Exactitude	49% (9,7%)	35% (7,4%)
Exactitude contiguë	79,6% (5,4%)	65,4% (7,4%)
RMSE	1,27	1,92
Exactitude	0,9	1,15

Tableau 5.34 – Détails sur les deux modèles finalement retenus pour les échelles à 6 et à 9 classes.

même un professeur de FLE sans expérience a peu de chance de commettre une telle erreur ? C'est cette quantité d'erreurs graves que reflète l'inverse de l'exactitude contiguë, lequelle atteint tout de même $21,4\%$ et $34,6\%$ pour notre formule. Cela signifie qu'environ un texte sur quatre est classé plus mal qu'à un niveau de distance de sa difficulté réelle. Idéalement, cette valeur devrait évidemment être proche de 0%, mais cela ne semble actuellement pas possible au vu de la difficulté de la tâche.

Afin d'éclairer un peu la répartition de ce quart de textes mal classés, un exemple de distribution des prédictions est reproduit pour chacun de deux modèles, dans les Tableaux 5.35 et 5.37. Il s'agit de la matrice de confusion de l'un des dix échantillons de tests utilisés pour évaluer le modèle, qui a été choisi de sorte que ses valeurs d'acc et d'$acc - cont$ soient les plus proches possible des valeurs moyennes obtenues sur les 10 échantillons.

	Prédiction					
Réel	1	2	3	4	5	6
1	**4**	**1**	0	0	0	0
2	0	**4**	**1**	1	1	0
3	0	**1**	**3**	0	0	2
4	0	2	**1**	**3**	**1**	0
5	0	0	1	0	**4**	**2**
6	0	0	0	1	0	**5**

Tableau 5.35 – Exemple de matrice de confusion pour le modèle SVM à 6 classes. Elle correspond à une acc de 60% et une $acc - cont$ de 79%. Les comptes associés à des observations correctes, sous l'hypothèse de l'exactitude contiguë, sont graisés, tandis que les erreurs graves sont indiquées en rouge.

Bien sûr, ces deux exemples ne peuvent prétendre être représentatifs de l'ensemble des prédictions des deux modèles, mais ils en illustrent toutefois bien certaines caractéristiques. En ce qui concerne le problème à 6 classes, on remarque que la majorité (79%) des prédictions sont bien situées sur la diagonale de la matrice ou sur une colonne adjacente. Néanmoins,

cette formule a tout de même classé un texte A2 comme C1 et deux textes B1 comme C2, ce qui représente de graves erreurs. Ces cas restent heureusement rares et apparaissent peu dans les niveaux débutants ou très avancés. En effet, une caractéristique de nos modèles, déjà mise en évidence précédemment (François, 2009a), est qu'ils se comportent mieux sur les classes situées aux deux extrémités du continuum de l'échelle de difficulté, à savoir A1, C1 et C2, comme le confirment les taux d'exactitude contiguë par niveau (*cf.* Tableau 5.36).

	A1	A2	B1	B2	C1	C2
Exact. cont.	100%	71%	67%	71%	86%	83%

Tableau 5.36 – Taux d'exactitude contiguë par niveau pour le modèle SVM à 6 classes, calculée sur l'un des 10 échantillons de test dont l'$acc = 60\%$ *et l'*$acc - cont = 79\%$.

Quant au modèle à 9 classes, sa matrice de confusion présente logiquement plus d'erreurs graves, puisque l'$acc - cont$ est inférieure. Il lui arrive ainsi de classer un texte B2 comme étant A1 ou un fragment A2 comme B2. Néanmoins, il faut souligner que ces performances inférieures par rapport à la version à 6 classes s'expliquent essentiellement en raison du raffinement de l'échelle. Celui-ci conduit à considérer comme une faute grave le fait de classer un texte A2+ comme A1+, alors que cette même prédiction est perçue comme acceptable dans le modèle à 6 classes. Nous avons d'ailleurs recalculé l'$acc - cont$ de cette matrice de confusion en fusionnant les prédictions sur 6 niveaux et avons obtenu une exactitude contiguë de 76%, un résultat fort proche des performances du classifieur à 6 classes. Nous aurions donc tendance à attribuer les dissimilarités de performances entre les deux modèles à la quantité différente de classes plutôt qu'à une infériorité des classifieurs.

	Prédiction								
Réel	1	2	3	4	5	6	7	8	9
1	**5**	**1**	0	0	0	0	0	0	0
2	**1**	**4**	**2**	0	0	0	0	0	0
3	0	**1**	**2**	**1**	1	0	1	0	0
4	1	1	**1**	**3**	0	1	0	0	0
5	0	0	2	**2**	0	**1**	0	0	1
6	0	0	0	0	**2**	**1**	**1**	3	0
7	1	1	1	1	0	0	**1**	**1**	1
8	0	0	1	0	1	0	**1**	**1**	**2**
9	0	0	0	0	0	1	0	**1**	**5**

Tableau 5.37 – Exemple de matrice de confusion pour le modèle SVM à 9 classes. Elle correspond à une acc *de* 37% *et une* $acc - cont$ *de* $67,8\%$. *Les comptes associés à des observations correctes, sous l'hypothèse de l'exactitude contiguë, sont graisés, tandis que les erreurs graves sont indiquées en rouge.*

Par ailleurs, le même motif que la formule à 6 classes apparaît lorsque l'on consulte le taux d'exactitude contiguë par niveau (*cf.* Tableau 5.38). Les textes A1 et A1+ sont le mieux dis-

criminés, suivis par les C2 et les A2 et C1. Par contre, les prédictions se dégradent nettement pour l'ensemble des classes médianes : A2+, B1, B1+ et B2. Ce profil s'explique par une plus grande similitude des observations situées dans ces niveaux, comme l'illustre la Figure 5.3, en ce qui concerne les fréquences lexicales. Au contraire, les textes A1 et A1+ se différencient particulièrement de ceux des autres niveaux par leur simplicité et sont donc faciles à détecter.

	A1	A1+	A2	A2+	B1	B1+	B2	C1	C2
Exact. cont.	100%	100%	67%	57%	50%	57%	29%	67%	86%

Tableau 5.38 – Taux d'exactitude contiguë par niveau pour le modèle SVM à 6 classes, calculée sur l'un des 10 échantillons de test dont l'$acc = 60\%$ *et l'*$acc - cont = 79\%$*.*

En résumé, même si ces deux formules se comportent nettement mieux que la « baseline » classique, la qualité de leurs prédictions apparaît encore un peu limite pour être appliquée de manière automatique dans le domaine de l'enseignement du FLE. Il reste quelques erreurs trop graves pour être tolérées au sein d'un logiciel d'exercices automatisés, qui générerait des items adaptatifs à partir de matériaux échelonnés par nos formules de lisibilité. Ces erreurs portent surtout sur les niveaux médians, qui comprennent des textes parfois assez similaires. Nous avions envisagé la possibilité qu'augmenter le nombre de variables au sein de modèles permette de découvrir des dimensions textuelles capables de mieux différencier ces observations problématiques. Or, malgré l'insertion de 46 prédicteurs, traitant de caractéristiques textuelles diverses, cela n'a pas été le cas. Cette question reste donc ouverte pour de prochaines tentatives d'amélioration de ce travail.

Comparaison avec la littérature

Au terme de la description des modèles retenus, le lecteur pourrait avoir l'impression que les performances obtenues semblent insatisfaisantes et douter qu'elles représentent une réelle avancée dans le domaine. Ce serait sans compter avec la difficulté de la tâche. En effet, les modèles décrits précédemment, non contents de classer trois fois mieux qu'un modèle basé sur le hasard, atteignent des performances consistantes par rapport à la littérature en lisibilité computationnelle.

Nous avons en effet repris dans le Tableau 5.39 les résultats rapportés par les principales études récentes. Ceux-ci sont évidemment difficilement comparables, étant donné qu'ils varient en fonction du corpus, du type de classifieur et, surtout, du nombre de niveaux de difficulté utilisés[32]. En effet, alors que le hasard permet d'atteindre 25% d'exactitude et

32. La procédure de comparaison habituelle en apprentissage automatisé aurait consisté à implémenter ces méthodes afin de les appliquer à nos propres données. Cependant, cette approche n'était pas réalisable dans

$62,5\%$ d'$acc - cont$ sur un modèle à quatre classes tel que celui de Schwarm et Ostendorf (2005), ses performances tombent à 11% d'exactitude et 31% d'$acc - cont$ sur un modèle à neuf classes. Il est donc clair que ces deux types de problèmes ne peuvent être directement comparés.

Étude	♯ classes	Langue	Type de modèle	Exac.	Exac. cont.	R	RMSE
Si et Callan (2001)	3	ENG.	Comb. linéaire	$75,4\%$	/	/	/
Collins-Thompson et Callan (2004a)	6	ENG.	Unig. lissé	/	/	$0,64$	/
Collins-Thompson et Callan (2004a)	12	ENG.	Unig. lissé	/	/	$0,79$	/
Collins-Thompson et Callan (2004a)	5	FR.	Unig. lissé	/	/	$0,64$	/
Schwarm et Ostendorf (2005)	4	ENG.	SVM	/	79% à $94,5\%$	/	/
Heilman *et al.* (2007)	12	ENG.	Unig. + KNN	/	/	$0,72$	$2,17$
Heilman *et al.* (2007)	4	ENG. (L2)	Unig. + KNN	/	/	$0,81$	$0,66$
Heilman *et al.* (2008)	12	ENG.	RLM	/	45%	$0,58$	$2,94$
Heilman *et al.* (2008)	12	ENG.	RLO	/	52%	$0,77$	$2,24$
Pitler et Nenkova (2008)	5	ENG.	Rég. linéaire	/	/	$0,78$	/
François (2009b)	6	FR. (L2)	RLM	41%	71%	$0,62$	/
François (2009b)	9	FR. (L2)	RLM	32%	63%	$0,72$	$2,24$
Feng *et al.* (2009)	4	ENG.	Rég. linéaire	/	/	$-0,34$	$0,57$
Feng *et al.* (2010)	4	ENG.	SVM	70%	/	/	/
Kate *et al.* (2010)	5	ENG.	Bagging	/	/	$0,82$	/
Modèle à 6 classes	6	FR. (L2)	SVM	49%	80%	$0,73$	$1,23$
Modèle à 9 classes	9	FR. (L2)	SVM	35%	65%	$0,74$	$1,92$

Tableau 5.39 – Comparaison des performances de notre modèle par rapport à celles de la littérature.

Pourtant, il apparaît que notre formule à 6 niveaux atteint une $acc - cont$ moyenne supérieure à celle que le modèle de Schwarm et Ostendorf (2005), comportant seulement 4 classes, obtient sur le *grade* 5 ($acc - cont = 79\%$) et pas si éloignée de celle atteinte sur le *grade* 4 ($acc - cont = 87\%$). Une autre manière de comparer ces deux approches est de regarder le gain d'$acc - cont$ par rapport au hasard. Chez Schwarm et Ostendorf (2005), il est respectivement de 29% pour le *grade* 5 et de $24,5$ pour le *grade* 4, tandis qu'il atteint 36% en moyenne pour notre modèle à 6 classes.

De même, par rapport à des problèmes à 5 classes, le R de notre modèle n'est que de $0,05$ inférieur à celui de (Pitler et Nenkova, 2008) et de $0,09$ par rapport à Kate *et al.* (2010), mais dépasse de $0,09$ le modèle de Collins-Thompson et Callan (2004a) pour le français.

Quant au modèle à 9 classes, il est évidemment plus précis que ceux à 12 classes de Heilman *et al.* (2008) ($acc-cont$ de 65% contre 52%). Il n'est pas évident de déterminer si ce gain de 13% s'explique uniquement par la différence de complexité du problème ou s'il correspond à une amélioration de la formule. Néanmoins, la différence entre les deux « baselines » basée sur le hasard est de 9% en $acc-cont$, ce qui reste inférieur à ces 13% entre les deux modèles.

notre cas, pour différentes raisons. Tout d'abord, la plupart de ces techniques utilisent des données d'entraînement anglophones dont nous ne disposions pas et qui ne sont surtout pas adaptées à nos objectifs. De plus, certains outils de paramétrisation requis sont uniquement disponibles pour l'anglais, en particulier en ce qui concerne les aspects sémantiques. Enfin, la plupart de ces formules utilisent les *grades* américains comme échelle, ce qui aurait engendré des problèmes de comparaison avec notre propre échelle.

Quoi que l'on en conclût, il est clair que notre classifieur se comporte également de manière tout à fait satisfaisante par rapport à ceux de Heilman *et al.* (2008).

En conclusion de ces différentes comparaisons, nous pouvons affirmer que les deux formules définies dans le cadre de cette étude sont tout à fait compétitives par rapport à la littérature en lisibilité computationnelle. Certes, si on la compare avec la seule formule de type classique publiée récemment pour le FLE (Uitdenbogerd, 2005), laquelle n'utilise que deux prédicteurs, il est vrai que nos R sont inférieurs au $0,87$ rapportés par cette auteur. Toutefois, il faut souligner que nos R ont été calculés sur 400 à 600 textes de nature variée et non sur 10 incipits de romans. De plus, ils sont issus d'un processus de validation croisée, ce qui ne semble pas être le cas des mesures rapportées par Uitdenbogerd (2005). Ces deux limitations méthodologiques laissent penser que les performances de nos modèles resteront plus stables lorsque ceux-ci seront confrontés à des textes inconnus, et en particulier à d'autres genres de textes que des romans.

5.5 Conclusion

Ce chapitre a décrit diverses expériences conduites en vue de répondre à deux objectifs : (1) évaluer l'apport des technologies de TAL à la lisibilité et ainsi évaluer la pertinence de ce nouveau courant que nous avons nommé la lisibilité computationnelle et (2) concevoir une nouvelle formule pour le FLE, qui s'intègre dans ce paradigme et soit capable de rivaliser avec les meilleures réalisations actuelles, aussi bien sur le français que sur l'anglais. Ces deux objectifs nous semblent avoir été remplis au terme de ce chapitre.

Tout d'abord, différents faits ont confirmé que l'emploi de technologies de TAL peut conduire à des avancées en lisibilité. Ainsi, l'analyse bivariée des prédicteurs a mis en évidence l'intérêt de certaines variables typiquement TAL, telles que celles basées sur la LSA ou sur la détection de certaines catégories grammaticales. Cependant, elle a aussi montré que l'emploi du TAL ne conduit pas nécessairement à des variables plus efficaces qu'une approche classique. Non seulement, le meilleur prédicteur s'est révélé être le pourcentage d'absents d'une liste issue d'un manuel de FLE, mais certaines des variables typiquement TAL, telles que les N-grammes et les UPs extraites automatiquement, n'ont pas débouché sur des résultats satisfaisants. Cette première analyse a donc produit un tableau mitigé des apports de cette discipline à la lisibilité, qui s'explique notamment par les approximations et erreurs que sous-tend l'automatisation de l'analyse du langage.

Dans la seconde partie de ce chapitre, l'intérêt du TAL est apparu plus nettement : il a été possible de concevoir une formule en moyenne 15% plus exacte qu'un modèle de type classique appliqué aux mêmes données. On pourrait évidemment épiloguer sur le fait que ce

sont surtout les techniques d'apprentissage automatisé qui sont à l'origine de la plus grande part du gain de performance. La confirmation de cette hypothèse demanderait de comparer notre approche, basée sur des variables linguistiques réfléchies, avec une approche brute, où les textes seraient représentés sous la forme de vecteurs de mots et où la seule source de connaissance du modèle serait obtenue lors de son entraînement sur les données. Toutefois, l'existence d'une interaction entre l'efficacité des divers types d'algorithmes statistiques et la quantité de prédicteurs qu'ils combinent se positionne déjà en défaveur de cette idée. Nous pensons plutôt que l'emploi des techniques d'apprentissage automatisé a rendu possible une utilisation optimale de l'information contenue dans les 47 prédicteurs retenus, dont certains relèvent de technologies TAL.

En tout cas, les apports de ces deux disciplines ont permis de concevoir deux nouvelles formules de lisibilité pour le FLE, un contexte pour lequel les réalisations en lisibilité se comptent sur les doigts de la main. De plus, il s'agit des premières formules de FLE issues du paradigme de la lisibilité computationnelle, ainsi que les premières à employer l'échelle de difficulté du CECR, ce qui les rend particulièrement attractives pour le monde de l'enseignement des langues étrangères. Elles présentent respectivement une exactitude de 49%, pour le modèle à 6 niveaux, et de 35% pour le modèle à 9 niveaux. Cela peut paraître faible, mais nous avons montré que ces résultats se situaient dans la droite ligne des réalisations actuelles et que la tâche de prédire automatiquement la difficulté de textes reste fort ardue pour les techniques actuelles. En effet, il est douteux que des formules commettant une erreur moyenne de respectivement $0,9$ ou $1,15$ niveau soit suffisamment matures pour des emplois complètement automatisés au sein d'applications d'ALAO. Elles autorisent néanmoins certaines perspectives sur lesquelles nous reviendrons dans la conclusion générale de cette étude. Par ailleurs, il est possible que, comme l'avaient montré sur leur propres données Kate *et al.* (2010), ce résultat automatique ne soit pas nécessairement moins bon que celui que produiraient des humains sans aucune expertise. On peut dès lors imaginer des contextes d'application où la capacité de discrimination de nos formules se révélerait déjà fort utile (*cf.* Conclusion).

Enfin, au-delà des deux objectifs précités, les expérimentations réalisées dans ce chapitre ont également permis de prendre position par rapport à une question centrale du domaine qui concerne le nombre optimal de prédicteurs. Celle-ci voit s'opposer la parcimonie classique au point de vue d'auteurs plus récents (Bormuth, 1966; Henry, 1975; Schwarm et Ostendorf, 2005), qui estiment nécessaire de combiner plus de deux ou trois variables afin d'obtenir une formule robuste et efficace. Sur ce plan, nos résultats se situent clairement dans la lignée de cette seconde position : (1) les formules les plus efficaces obtenues utilisent rien moins que 46 prédicteurs ; (2) nos expérimentations avec deux, quatre, huit et 46 variables ont montré que plus leur nombre augmentait, plus les performances des formules tendent à

faire de même et (3) le modèle obtenu automatiquement pour la RLO, assimilable à une adaptation moderne de l'idéal classique et supposé optimal par l'algorithme de sélection automatique de variables, a été surpassé par l'ensemble des autres modèles automatiques. Nous en concluons que la difficulté constitue bien un phénomène complexe aux multiples facettes, qui ne saurait être réduit à une ou deux dimensions textuelles, du moins dans l'état actuel de nos connaissances.

Conclusion

Nous avons vu que la lisibilité comptait une riche tradition d'études, bien que celles-ci soient surtout centrées sur l'anglais L1. Dans cette étude, nous avons voulu traiter d'une problématique nettement moins abordée : la lisibilité du FLE. Celle-ci a été envisagée en fonction d'une double problématique : déterminer la pertinence des apports du TAL dans le contexte spécifique de la lisibilité du FLE et proposer une nouvelle formule de lisibilité qui profite le cas échéant de ces apports. Au-delà de ces deux aspects, ce travail a aussi abouti sur plusieurs résultats annexes intéressants, parmi lesquels on peut mettre en évidence une nouvelle synthèse du domaine de la lisibilité prenant en compte les travaux récents ; un ensemble de nouveaux indices de la difficulté, articulés au sein de l'examen critique des facteurs de lisibilité utilisés dans la littérature, ou encore la description d'une technique de récolte de textes qui permette d'obtenir une formule compatible avec le CECR. Ces différents apports sont discutés en détail par la suite.

Ainsi, la description, aussi exhaustive que possible, des nombreux travaux en lisibilité L1 et L2 a représenté un travail de synthèse considérable qui n'avait plus été accompli depuis de nombreuses années, en particulier en ce qui concerne le français. Nous espérons qu'il pourra servir de référence pour les prochaines études sur la question, que nous souhaitons nombreuses [33]. Nous avons aussi proposé, à partir de cet état de l'art, une vision originale de la discipline dans laquelle l'opposition entre les études classiques et structuro-cognitives est figurée selon un schéma thèse-antithèse, dont la synthèse serait réalisée au sein d'un nouveau paradigme que nous avons identifié sous le terme de lisibilité computationnelle.

En effet, l'un des aspects que ce paradigme récent cherche à développer sont les dimensions textuelles prises en compte par les formules de lisibilité. Nous avons vu que les formules classiques reposent essentiellement sur des caractéristiques lexico-syntaxiques, tandis que les structuro-cognivistes concentrent surtout leurs efforts sur le niveau sémantique, au risque d'obtenir des variables difficiles à paramétriser automatiquement. La lisibilité computationnelle, parce qu'elle vise à une mécanisation de ces modèles sémantiques complexes, permet

33. Nous avons d'ailleurs déjà proposé une version synthétique de cet historique dans un article récent (François, 2011), lequel se focalise plus particulièrement sur les études pour le français.

quant à elle d'intégrer ces deux types d'information au sein d'un même modèle et envisage dès lors tout type de variables. Endossant cette vision, le chapitre 3 a non seulement proposé un inventaire structuré des prédicteurs utilisés dans la littérature, mais a aussi permis d'identifier de nouvelles variables, définies à partir de données expérimentales issues de la psychologie de la lecture. Parmi ces nouveaux facteurs, citons le nombre de voisins orthographiques, la structure des unités polylexicales, le TTR normalisé, les étapes d'acquisition ou encore le nombre de sens partagés par une même forme lexicale.

Qu'il s'agisse des variables déjà reconnues dans le domaine ou de celles que nous avons proposées, leur présentation a été accompagnée à chaque fois d'une discussion critique. Celle-ci mêle des arguments théoriques, tirés de la psychologie cognitive et visant à déterminer s'il existe bel et bien un effet de chacune de ces variables sur le processus de lecture, avec des résultats d'expériences précédentes en lisibilité qui concluent tantôt en faveur, tantôt en défaveur de ces indices de la difficulté. Cet examen critique constitue probablement la contribution la plus intéressante de ce chapitre, en ce sens qu'il effectue un tri entre les facteurs qui ne sont que de simples indices de la difficulté textuelle et ceux qui constituent des causes de la difficulté, ouvrant ainsi la porte à des travaux portant sur la simplification textuelle.

Par ailleurs, il faut souligner que si cet examen critique a permis de valider l'intérêt de nombreuses variables, y compris celles issues de la mouvance structuro-cognitive, la synthèse des résultats issus de la littérature a montré que l'efficacité de certaines de ces dernières pouvait faire débat. En effet, nous avons vu que ces variables étaient fort complexes à opérationnaliser et que lorsqu'un traitement automatisé leur était appliqué, celui-ci avait tendance à entraîner des erreurs en cascades dans les différents modules TAL mobilisés, qui en faisaient de médiocres prédicteurs. Par conséquent, la question des apports du TAL à la lisibilité n'a pu être tranchée seulement sur la base de la littérature.

C'est pourquoi nous avons entrepris d'effectuer nos propres expérimentations. Pour ce faire, nous avons implémenté un sous-ensemble des variables décrites au Chapitre 3 et en avons testé l'efficacité sur un corpus de textes spécifiques au FLE. La constitution de ce corpus a été rapportée dans le Chapitre 4 et s'appuie sur une comparaison critique des différentes stratégies disponibles pour annoter un corpus de textes en fonction de leur difficulté. La technique retenue pour cette étude est l'extraction de textes depuis des manuels de FLE dont le niveau a été défini manuellement par des professionnels de la didactique du FLE. Elle repose sur l'hypothèse qu'un texte issu d'un manuel positionné par ses auteurs au niveau X peut se voir attribuer ce niveau X. Comme l'ensemble des manuels utilisés sont étalonnés selon l'échelle du CECR, il a été possible d'exprimer les prédictions de nos modèles selon ce standard. Il s'agit d'un avantage déterminant de notre approche, puisque cette échelle

n'a encore jamais été employée en lisibilité des langues étrangères, alors qu'elle représente aujourd'hui la norme dans le milieu de l'enseignement des langues étrangères en Europe.

Nous avons montré que cette hypothèse de l'équivalence de niveau entre un texte et son manuel est valide, mais qu'elle comporte certaines faiblesses ayant rendu nécessaire le nettoyage du corpus. Huit corpus différents ont ainsi résulté des différentes opérations de nettoyage. Certains d'entre eux ont vu les manuels comportant des textes annotés de manière trop incohérente par rapport aux autres textes être supprimés, tandis que d'autres ont simplement vu leurs textes de niveau C2 être divisés en fragments plus réduits, afin d'augmenter le nombre d'observations disponibles pour l'entraînement des modèles statistiques. D'après les résultats rapportés au Chapitre 5, il est probable que cette opération ait contribué à améliorer les résultats de nos modèles, étant donné que les performances du modèle RLM avec sélection de variables automatisées se sont révélées supérieures à notre modèle de 2009 (François, 2009b), très similaire par ailleurs, mais entraîné sur un corpus non nettoyé.

Malgré l'efficacité de ce nettoyage, la question de l'optimalité du critère utilisé pour annoter les textes de notre corpus de référence reste discutable. Le critère que nous avons employé est le jugement d'expert, lequel connaît actuellement un grand succès dans le domaine. Néanmoins, bien qu'il soit certainement l'un des plus pratiques, il n'est pas évident qu'il soit le plus adapté pour cette étape de constitution des corpus d'entraînement. En effet, van Oosten *et al.* (2011) a également rapporté une certaine hétérogénéité entre les jugements d'experts. Notre choix de critère ayant été opéré sur la base d'études théoriques, rien ne nous dit qu'un autre critère, tel que le taux de rappel, le temps de lecture ou le test de closure, n'aurait pas produit un corpus annoté de manière moins hétérogène, permettant à nos formules d'atteindre des résultats plus élevés encore.

Malheureusement, la problématique de l'annotation du corpus est peu posée actuellement en lisibilité computationnelle et nous manquons de données de comparaison pour poser un jugement critique sur l'optimalité de notre approche. De plus, les autres techniques ne sont pas nécessairement mieux considérées. Ainsi, le test de closure, qui représente un autre outil d'annotation classique en lisibilité, a également fait l'objet de critiques dans le contexte L2. Greenfield (2004) laisse ainsi penser que les apprenants de L2 pourraient être plus habitués que des natifs à construire une représentation sémantique d'un texte malgré un décodage imparfait et que leur compréhension serait dès lors supérieure à ce que révèlent leurs résultats à ce type de tests. Toutefois, à notre connaissance, cette hypothèse n'a pas été investiguée plus en détail. En l'occurrence, il nous semble que la question des problèmes d'annotation du corpus de référence mériterait d'être davantage explorée dans des études postérieures, afin d'évaluer la part d'erreurs de classification qu'elle explique.

Une fois cette étape de constitution terminée, nous avons réalisé une série d'expériences,

rapportées dans le Chapitre 5. Celles-ci ont permis de répondre aux deux problématiques soulevées dans l'introduction de cette thèse. D'une part, il s'est avéré que les formules exploitant un certain nombre de prédicteurs issus des quatre grandes familles délimitées au Chapitre 3 sont significativement plus exactes qu'une adaptation moderne des formules classiques reposant sur des SVM et ne recourant qu'à deux variables lexico-syntaxiques. Il est donc clair que l'emploi de techniques d'apprentissage automatisé permet de définir des modèles plus complexes et plus efficaces, sans que cela ne soit une charge supplémentaire pour l'utilisateur, puisque la paramétrisation des variables est automatique.

Par contre, en ce qui concerne le TAL à proprement parler, les résultats restent plus mitigés. En effet, plusieurs des variables basées sur des modèles spécifiques à cette discipline, tels les n-grammes, l'extraction d'unités polylexicales ou l'annotation automatique en catégories grammaticales, se sont révélées peu intéressantes aussi bien lors de l'analyse univariée que lors de l'étape de modélisation. À ce propos, nos observations ont suscité plusieurs interrogations qui gagneraient à être envisagées dans une prochaine étude. Citons entre autres la divergence d'efficacité entre nos modèles n-grammes et les résultats rapportés dans la littérature ; la possibilité que les variables d'UPs deviennent plus discriminantes si l'on ne considérait que les expressions sémantiquement opaques, ou encore le fait que la cohérence des textes mesurés via la LSA conclut que, sur nos données, plus un texte est cohérent, plus il est difficile à lire.

Nos analyses univariées ont toutefois mis en évidence la qualité de deux variables typiquement TAL : les modèles unigrammes et le niveau de cohérence des textes mesurés à l'aide de la LSA. Si l'intérêt des premiers a été confirmé lors de la seconde phase d'analyse, qui prenait en compte le recouvrement informationnel entre variables, il est par contre apparu que les seconds n'ont pas été retenus dans nos modèles avec sélection automatisée des variables. De plus, il est même arrivé que la variable **avLocalLsa_Lem**, pourtant si efficace seule, soit considérée comme colinéaire, ce qui signifie qu'elle n'apportait aucune information nouvelle par rapport à celle déjà présente dans les autres variables. Par conséquent, même s'il semble que l'association entre le TAL et la lisibilité du FLE puisse être féconde, ainsi que l'avaient déjà montré précédemment certains chercheurs sur l'anglais L1 Collins-Thompson et Callan (2004a); Heilman *et al.* (2007); Schwarm et Ostendorf (2005); Si et Callan (2001), l'apport des techniques d'apprentissage automatisé pourrait bien être plus crucial encore. Nos expériences n'ont toutefois pas distingué clairement ces deux apports, ce qui serait nécessaire afin de prendre une position tranchée sur la question.

Nous avions également émis une autre hypothèse au cours des Chapitres 3 et 5, à savoir que les améliorations les plus sensibles de nos modèles proviendraient de l'association de variables issues des quatre grandes familles que nous avons distinguées. Or, les performances

des modèles experts, s'ils sont apparus significativement supérieurs à la « baseline », n'ont pas conduit à un gain aussi large qu'espéré. Au contraire, il est apparu que c'est surtout l'addition d'un plus grand nombre de variables au sein de modèles suffisamment complexes pour tirer parti de cette richesse informationnelle qui a engendré les meilleures performances. Nous aurions donc tendance à conclure à un recoupement important entre les niveaux sémantique et lexico-syntaxique, pourtant farouchement opposés dans la littérature des années 80, comme le montre le cas de **avLocalLsa_Lem**, qui vient d'être discuté. Ce résultat est d'ailleurs cohérent avec les expériences de Kemper (1983), laquelle n'avait pas obtenu d'amélioration significative de sa formule en combinant un facteur traditionnel avec trois facteurs cognitifs.

Enfin, le principal résultat de cette thèse est évidemment la nouvelle formule de lisibilité auquel elle a abouti. Celle-ci a été déclinée selon deux échelles de difficulté : l'une à 6 niveaux et la seconde à 9 niveaux. La première atteint une exactitude contiguë de 80%, tandis que la seconde obtient 65%. Toutes deux apparaissent parfaitement en phase avec les performances des modèles les plus récents pour l'anglais et, bien qu'elles présentent un R inférieur à la formule de Uitdenbogerd (2005), plusieurs éléments laissent supposer que leur capacité de généralisation sur différents types de texte est supérieure à celle de cette dernière.

L'existence d'un tel outil, qui ne devrait pas tarder à être rendu public, ouvre évidemment de nombreuses perspectives au sein d'applications d'ALAO. Nous avons déjà cité dans l'introduction l'exemple de la génération automatique d'exercice, au sein de laquelle les phrases servant de support sont aujourd'hui étalonnées manuellement, quand elles le sont. L'insertion d'une formule de lisibilité au sein d'une telle application rendrait possible le contrôle de la difficulté des matériaux servant de support à la tâche. Le seul problème vis-à-vis de cette perspective est, comme nous l'avons déjà souligné, le fait que le taux d'exactitude contiguë actuel des formules indique l'existence de quelques erreurs graves, susceptibles de décourager l'apprenant dans sa pratique de la langue.

Par contre, les performances de nos modèles en autorisent un usage semi-automatisé, dans lequel la formule opère une annotation d'un ensemble de textes, avant que celle-ci ne soit vérifiée par un humain. Une application correspondant parfaitement à cette description est la recherche de textes à visée pédagogique sur le web. En effet, un modèle de difficulté permettrait de récupérer un sous-ensemble de documents d'une difficulté cible sur le web, parmi lequel l'utilisateur aurait plus de chance de trouver rapidement son bonheur. Nous avons signalé que plusieurs outils de ce type existaient pour l'anglais (Collins-Thompson et Callan, 2004b; Miltsakaki, 2009; Uitdenbogerd, 2010), mais rien de tel n'est encore disponible pour le FLE. Pourtant, diverses discussions avec des professeurs de FLE nous ont convaincu que

ce type de logiciels répondait à un besoin réel de la profession.

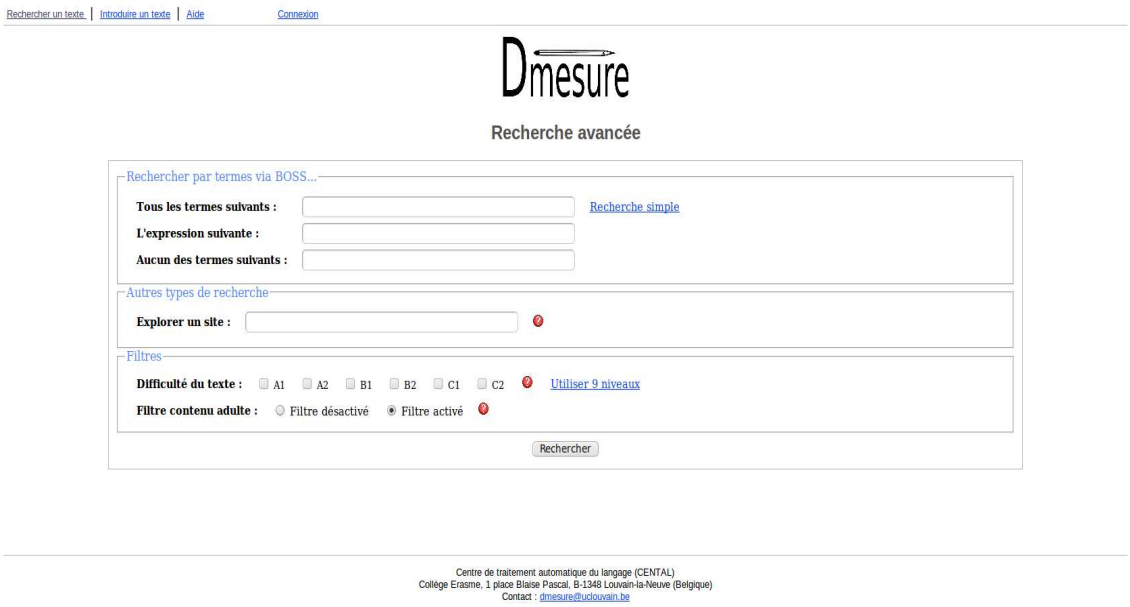
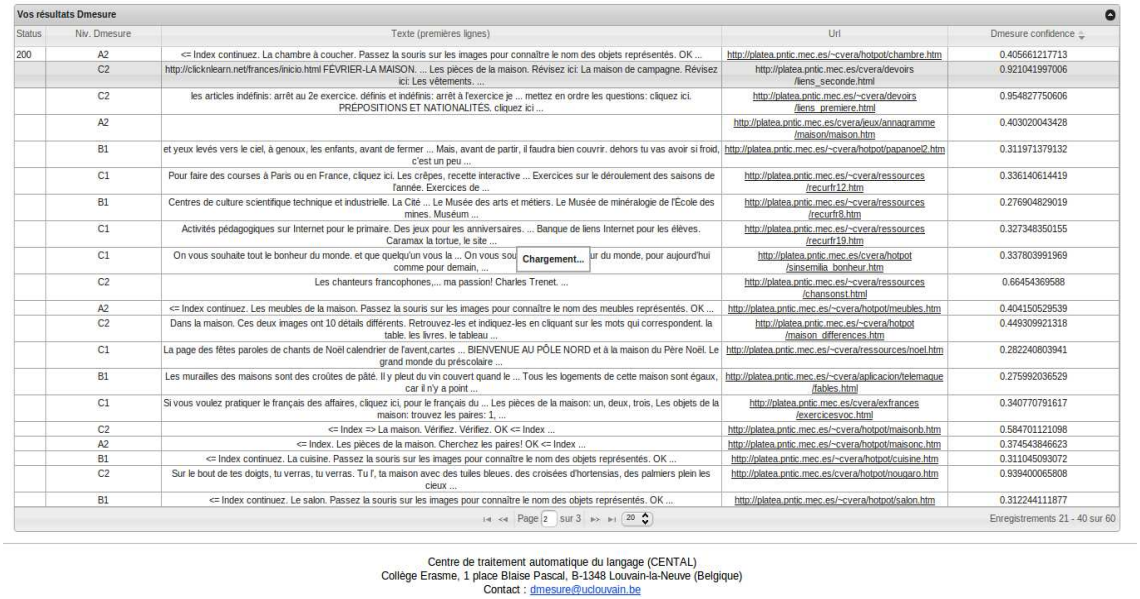


Figure 5.14 – Interface de recherche de Dmesure.

C'est pourquoi nous avons, parallèlement à cette thèse, implémenté le prototype d'une plateforme en ligne, appelée « Dmesure », qui illustre bien les possibilités d'applications ouvertes par notre travail [34]. Elle vise à offrir trois services différents. Tout d'abord, elle permet d'ores et déjà d'introduire un texte via une interface web et d'obtenir en retour une estimation de son niveau de difficulté. Ensuite, elle propose une interface épurée de recherche de documents en ligne, représentée à la Figure 5.14, à ceci près qu'il est possible de spécifier un ou plusieurs niveaux de difficulté pour les textes à récupérer. Sur cette capture d'écran, on remarque aussi la possibilité pour le professeur d'explorer un site unique, qu'il s'agisse de son site ressource pour le FLE ou de son site de presse préféré. La Figure 5.15 présente un exemple de résultats pour la requête « maison » sur un site de FLE. Concernant cette tâche de recherche, notre évaluation des performances actuelles de Dmesure sur cette tâche a montré que la plateforme tend à surestimer le niveau réel des textes sous l'effet du bruit présent sur les pages web où les textes sont récoltés. Il convient donc de concevoir un extracteur de textes efficace, capable d'éliminer ce bruit.

En plus de ces deux services, Dmesure présente également une dimension collaborative, apparentée au « crowd-sourcing », qui vise à améliorer les performances du modèle actuel. Pour ce faire, il offre à l'utilisateur préalablement identifié la possibilité d'accéder à l'historique des textes (représenté à la Figure 5.16), que celui-ci a précédemment soumis au

34. Nous remercions à cette occasion Hubert Naets pour l'implémentation du web service REST permettant à l'interface « Dmesure » de communiquer avec notre formule de lisibilité. Des détails supplémentaires sur ce prototype peuvent être trouvés dans un article publié précédemment (François, 2011).



Vos résultats Dmesure

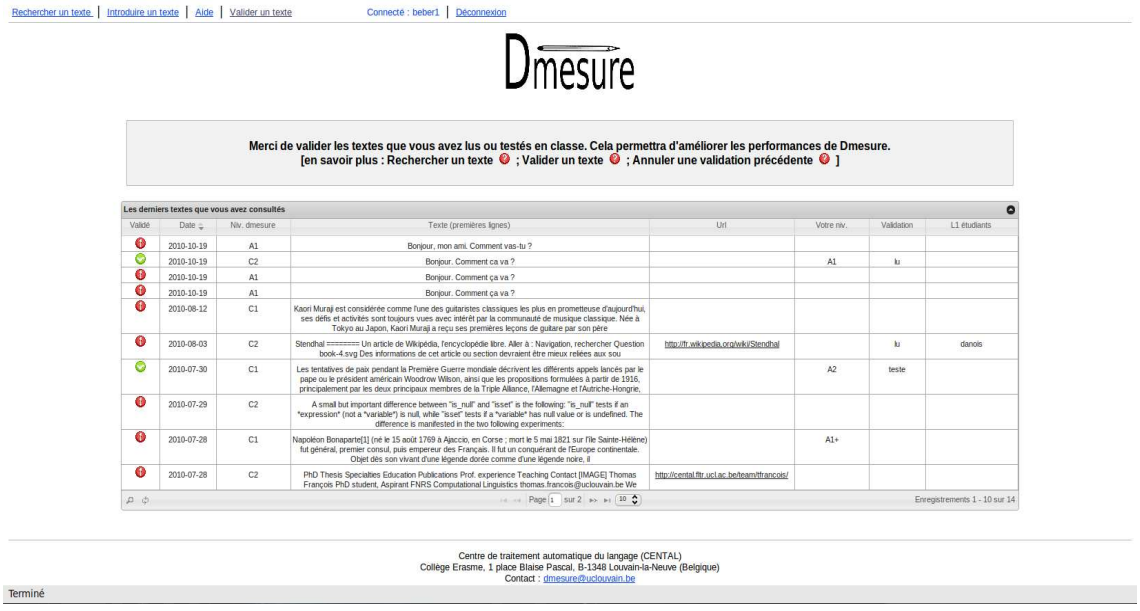
Status	Niv. Dmesure	Texte (premières lignes)	Url	Dmesure confidence
200	A2	<= Index continuez. La chambre à coucher. Passez la souris sur les images pour connaître le nom des objets représentés. OK ...	http://platea.pntic.mec.es/~cvera/hotpot/chambre.htm	0.405661217713
	C2	http://clicknlearn.net/frances/inicio.html FÉVRIER-LA MAISON. ... Les pièces de la maison. Révisez ici: La maison de campagne. Révisez ici: Les vêtements. ...	http://platea.pntic.mec.es/cvera/devoirs/liens_seconde.html	0.921041997006
	C2	les articles indéfinis: arrêt au 2e exercice. définis et indéfinis: arrêt à l'exercice je ... mettez en ordre les questions: cliquez ici. PREPOSITIONS ET NATIONALITÉS. cliquez ici ...	http://platea.pntic.mec.es/~cvera/devoirs/liens_premiere.html	0.954827750606
	A2		http://platea.pntic.mec.es/cvera/jeux/annagramme/maison/maison.htm	0.403020043428
	B1	et yeux levés vers le ciel, à genoux, les enfants, avant de fermer ... Mais, avant de partir, il faudra bien couvrir. dehors tu vas avoir si froid, c'est un peu ...	http://platea.pntic.mec.es/~cvera/hotpot/papanoel2.htm	0.311971379132
	C1	Pour faire des courses à Paris ou en France, cliquez ici. Les crêpes, recette interactive ... Exercices sur le déroulement des saisons de l'année. Exercices de ...	http://platea.pntic.mec.es/~cvera/ressources/recurfr12.htm	0.336140614419
	B1	Centres de culture scientifique technique et industrielle. La Cité ... Le Musée des arts et métiers. Le Musée de minéralogie de l'École des mines. Muséum ...	http://platea.pntic.mec.es/~cvera/ressources/recurfr8.htm	0.276904829019
	C1	Activités pédagogiques sur Internet pour le primaire. Des jeux pour les anniversaires. ... Banque de liens Internet pour les élèves. Caramax la tortue, le site ...	http://platea.pntic.mec.es/~cvera/ressources/recurfr19.htm	0.327348350155
	C1	On vous souhaite tout le bonheur du monde. et que quelqu'un vous la ... On vous sou Chargement... ur du monde, pour aujourd'hui comme pour demain, ...	http://platea.pntic.mec.es/cvera/hotpot/sinsemilia_bonheur.htm	0.337803991969
	C2	Les chanteurs francophones,... ma passion! Charles Trenet. ...	http://platea.pntic.mec.es/~cvera/ressources/chansonst.html	0.66454369588
	A2	<= Index continuez. Les meubles de la maison. Passez la souris sur les images pour connaître le nom des meubles représentés. OK ...	http://platea.pntic.mec.es/~cvera/hotpot/meubles.htm	0.404150529539
	C2	Dans la maison. Ces deux images ont 10 détails différents. Retrouvez-les et indiquez-les en cliquant sur les mots qui correspondent. la table. les livres. le tableau ...	http://platea.pntic.mec.es/~cvera/hotpot/maison_differences.htm	0.449309921318
	C1	La page des fêtes paroles de chants de Noël calendrier de l'avent,cartes ... BIENVENUE AU PÔLE NORD et à la maison du Père Noël. Le grand monde du préscolaire ...	http://platea.pntic.mec.es/~cvera/ressources/noel.htm	0.282240803941
	B1	Les murailles des maisons sont des croûtes de pâté. Il y pleut du vin couvert quand le ... Tous les logements de cette maison sont égaux, car il n'y a point ...	http://platea.pntic.mec.es/~cvera/aplicacion/telemaque/fables.html	0.275992036529
	C1	Si vous voulez pratiquer le français des affaires, cliquez ici, pour le français du ... Les pièces de la maison: un, deux, trois, Les objets de la maison: trouvez les paires: 1, ...	http://platea.pntic.mec.es/cvera/exfrances/exercicesvoc.html	0.340770791617
	C2	<= Index => La maison. Vérifiez. Vérifiez. OK <= Index ...	http://platea.pntic.mec.es/~cvera/hotpot/maisonb.htm	0.584701121098
	A2	<= Index. Les pièces de la maison. Cherchez les paires! OK <= Index ...	http://platea.pntic.mec.es/~cvera/hotpot/maisonc.htm	0.374543846623
	B1	<= Index continuez. La cuisine. Passez la souris sur les images pour connaître le nom des objets représentés. OK ...	http://platea.pntic.mec.es/~cvera/hotpot/cuisine.htm	0.311045093072
	C2	Sur le bout de tes doigts, tu verras, tu verras. Tu l', ta maison avec des tuiles bleues. des croisées d'hortensias, des palmiers plein les cieux ...	http://platea.pntic.mec.es/cvera/hotpot/nougaro.htm	0.939400065808
	B1	<= Index continuez. Le salon. Passez la souris sur les images pour connaître le nom des objets représentés. OK ...	http://platea.pntic.mec.es/~cvera/hotpot/salon.htm	0.312244111877

Page 2 sur 3 — 20 — Enregistrements 21 - 40 sur 60

Centre de traitement automatique du langage (CENTAL)
Collège Erasme, 1 place Blaise Pascal, B-1348 Louvain-la-Neuve (Belgique)
Contact : dmesure@uclouvain.be

Figure 5.15 – Exemple de résultats de Dmesure pour la requête « maison » envoyée à un site de FLE.

système ou qui ont été récupérés suite à l'une de ses requêtes. L'interface de cet historique offre différents services aux utilisateurs de Dmesure et ouvre des perspectives de recherches intéressantes.



Rechercher un texte | Introduire un texte | Aide | Valider un texte — Connecté : beber1 | Déconnexion

Dmesure

Merci de valider les textes que vous avez lus ou testés en classe. Cela permettra d'améliorer les performances de Dmesure.
[en savoir plus : Rechercher un texte ; Valider un texte ; Annuler une validation précédente]

Les derniers textes que vous avez consultés

Validé	Date	Niv. dmesure	Texte (premières lignes)	Url	Votre niv.	Validation	L1 étudiants
	2010-10-19	A1	Bonjour, mon ami. Comment vas-tu ?				
	2010-10-19	C2	Bonjour. Comment ca va ?		A1	lu	
	2010-10-19	A1	Bonjour. Comment ça va ?				
	2010-10-19	A1	Bonjour. Comment ça va ?				
	2010-08-12	C1	Kaori Muraji est considérée comme l'une des guitaristes classiques les plus en prometteuse d'aujourd'hui, ses défis et activités sont toujours vues avec intérêt par la communauté de musique classique. Née à Tokyo au Japon, Kaori Muraji a reçu ses premières leçons de guitare par son père				
	2010-08-03	C2	Stendhal ========= Un article de Wikipédia, l'encyclopédie libre. Aller à : Navigation, rechercher Question book-4.svg Des informations de cet article ou section devraient être mieux reliées aux sou	http://fr.wikipedia.org/wiki/Stendhal		lu	danois
	2010-07-30	C1	Les tentatives de paix pendant la Première Guerre mondiale décrivent les différents appels lancés par le pape ou le président américain Woodrow Wilson, ainsi que les propositions formulées à partir de 1916, principalement par les deux principaux membres de la Triple Alliance, l'Allemagne et l'Autriche-Hongrie,		A2	teste	
	2010-07-29	C2	A small but important difference between "is_null" and "isset" is the following: "is_null" tests if an *expression* (not a *variable*) is null, while "isset" tests if a *variable* has null value or is undefined. The difference is manifested in the two following experiments:				
	2010-07-28	C1	Napoléon Bonaparte[1] (né le 15 août 1769 à Ajaccio, en Corse ; mort le 5 mai 1821 sur l'île Sainte-Hélène) fut général, premier consul, puis empereur des Français. Il fut un conquérant de l'Europe continentale. Objet dès son vivant d'une légende dorée comme d'une légende noire, il		A1+		
	2010-07-28	C2	PhD Thesis Specialties Education Publications Prof. experience Teaching Contact [IMAGE] Thomas François PhD student, Aspirant FNRS Computational Linguistics thomas.francois@uclouvain.be We	http://cental.fltr.ucl.ac.be/team/tfrancois/			

Page 1 sur 2 — 10 — Enregistrements 1 - 10 sur 14

Centre de traitement automatique du langage (CENTAL)
Collège Erasme, 1 place Blaise Pascal, B-1348 Louvain-la-Neuve (Belgique)
Contact : dmesure@uclouvain.be

Terminé

Figure 5.16 – Exemple d'un historique d'utilisateur dans Dmesure.

Tout d'abord, les experts (professeurs, concepteurs de manuels, etc.) qui utilisent Dmesure pour évaluer les textes qu'ils emploient par ailleurs dans le cadre de leur fonction se voient offrir la possibilité de confirmer ou de corriger les prédictions de Dmesure en proposant leur propre évaluation des textes, basée sur une expérience pédagogique concrète. Ce geste

devrait permettre, à terme, une évaluation extrinsèque – et non pas uniquement intrinsèque – de nos formules, puisqu'il sera possible de comparer les performances de celles-ci à l'avis d'experts du FLE. D'autre part, les textes annotés par ces experts formeront également un nouveau corpus d'entraînement, dont on peut espérer qu'il sera moins hétérogène que celui dont nous disposions dans cette étude. Celui-ci pourrait alors servir à entraîner une nouvelle version de nos formules, mais aussi d'envisager des formules personnalisées, pour autant qu'un utilisateur (ou groupe d'utilisateurs) disposent de suffisamment de textes annotés par ses soins.

Par ailleurs, cet axe collaboratif pourrait être développé afin de permettre l'exploration d'autres champs de recherche encore. Dmesure pourrait ainsi se révéler un moyen précieux de collecte de données spécialisées, par exemple en ce qui concerne les interactions entre langues L1 et L2. En effet, l'interface de critique des textes par les professeurs permet d'ores et déjà de spécifier, pour chaque texte validé, la langue maternelle des apprenants avec lesquels ce texte a été utilisé en classe. Inversement, moyennant quelques arrangements, rien n'empêche que Dmesure devienne également une base de textes de FLE récoltés sur le web et classés d'après les niveaux du CECR à la fois par nos formules et par des utilisateurs experts. L'existence d'une telle base de données dans le milieu de l'enseignement du FLE serait, à notre avis, un outil particulièrement pratique pour les professionnels du secteur.

On voit donc que notre recherche offre de riches perspectives pratiques, mais également théoriques. En effet, comme c'est généralement le cas en recherche, une étude en appelle d'autres. Sachant que la lisibilité est un domaine peu parcouru dans le cadre francophone, il n'est pas évident que cela soit le cas à court terme, même si cette problématique semble à nouveau attirer l'attention de quelques jeunes chercheurs (Barbaresi, 2011; van Oosten *et al.*, 2011) ou d'entreprises telles que *Translated Labs* [35]. Au vu des nombreuses applications envisageables pour les outils de lisibilité, ne fût-ce que pour contrôler la difficulté des sorties textuelles de logiciels de traduction automatique ou de résumé automatique, il est certain qu'il s'agit d'un domaine appelé à se développer dans les prochaines années. Nous espérons que cette contribution pourra aider, autant que possible, au développement d'une discipline aussi riche par les problématiques qu'elle soulève que par les perspectives qu'elle ouvre.

35. Celle-ci propose un outil de lisibilité en ligne, qui classe les textes en facile, moyen et difficile. Il est accessible à l'adresse suivante `http://labs.translated.net/lisibilite-texte/`.

ANNEXE A

TABLES DES SIGLES ET ACRONYMES

AIC Akaike Information Criterion

ALAO apprentissage des langues assisté par ordinateur

ALS anglais comme langue seconde

ANOVA analyse de la variance

APS auto-présentation segmentée

ARI Automated Readability Index

ATOS Advantage-TASA Open Standard

BIA Bilingual Interactive Activation

BIC Bayesian Information Criterion

BNC British National Corpus

BOLD Blood Oxygen Level Dependant

CECR Cadre européen commun de référence pour les langues

DRC Dual Route Cascade

DRP Degrees of Reading Power

ELAN Early Left Anterior Negativity

ELF Easy Listening Formula

EM Expectation-Maximization

FLE français langue étrangère

FOS français sur objectifs spécifiques

GGT Grammaire Générative et Transformationnelle

HI Human Interest

HTML Hypertext Markup Language

IPsyn Index of Productive Syntax

IRMf Imagerie Fonctionnelle par Résonance Magnétique

kNN k plus proches voisins

L1 langue maternelle ou langue première

L2 langue seconde ou langue étrangère

LSA analyse sémantique latente

LVE langues vivantes étrangères

MA méthodologie active

MAE erreur moyenne absolue

MANOVA Multivariate ANalysis Of VAriance

MAO méthodologie audio-orale

MAs mesures d'association

MAV méthodologie audio-visuelle

MD méthodologie directe

MEG Magnéto-Encéphalographie

MSE erreur quadratique moyenne

MT méthodologie traditionnnelle

OCR Optical Character Recognition

PDP Parallel Distributed Processing

PEV Potentiels Électriques Évoqués

RLM régression logistique multinomiale

RLO régression logistique ordinale

PHAN PHrase ANalysis

RE Reading Ease

RBF Gaussian Radial Basis Function

RIDE Reading-Input Difficulty Estimate

RMSE erreur type

SEER Singer Eyeball Estimate of Readability

SGAV méthode structuro-globale audio-visuelle

SMOG Simple Measure of Gobbledygook

SVD décomposition en valeurs singulières

SVM séparateur à vaste marge

TAL Traitement Automatique du Langage

TASA Touchstone Applied Science Associates

TEI Text Encoding Initiative

TIC Technologies de l'Information et de la Communication

ThUS Thought Unit Sentences

TTR Type Token Ratio

TEP Tomographie par Émission de Positions

UE Union européenne

UP unité polylexicale

UPs unités polylexicales

USAÉ Unité Sémantique Abstraite Élémentaire

XCES Corpus Encoding Standard for XML

XML Extensible Markup Language

ANNEXE B

LISTE ALPHABÉTIQUE DES INDICES DE LISIBILITÉ UTILISÉS

75FFF	Probabilité équivalant au 75^{e} percentile de la distribution des fréquences des formes fléchies.
75FFFC	Probabilité équivalant au 75^{e} percentile de la distribution des fréquences des formes fléchies, uniquement pour les mots de contenu.
75FFFD	Probabilité équivalant au 75^{e} percentile de la distribution des fréquences des formes fléchies désambiguïsées.
75FFFDC	Probabilité équivalant au 75^{e} percentile de la distribution des fréquences des formes fléchies désambiguïsées, uniquement pour les mots de contenu.
75FFFDG	Probabilité équivalant au 75^{e} percentile de la distribution des fréquences des formes fléchies désambiguïsées, uniquement pour les mots grammaticaux.
75FFFDN	Probabilité équivalant au 75^{e} percentile de la distribution des fréquences des formes fléchies désambiguïsées, uniquement pour les noms.
75FFFG	Probabilité équivalant au 75^{e} percentile de la distribution des fréquences des formes fléchies, uniquement pour les mots grammaticaux.
75FFFN	Probabilité équivalant au 75^{e} percentile de la distribution des fréquences des formes fléchies, uniquement pour les noms.
75FL	Probabilité équivalant au 75^{e} percentile de la distribution des fréquences des lemmes.
75FLC	Probabilité équivalant au 75^{e} percentile de la distribution des fréquences du lemme, uniquement pour les mots de contenu.
75FLG	Probabilité équivalant au 75^{e} percentile de la distribution des fréquences du lemme, uniquement pour les mots grammaticaux.
75FLN	Probabilité équivalant au 75^{e} percentile de la distribution des fréquences du lemme, uniquement pour les noms.
90FFF	Probabilité équivalant au 90^{e} percentile de la distribution des fréquences des formes fléchies.

90FFFC	Probabilité équivalant au 90e percentile de la distribution des fréquences des formes fléchies, uniquement pour les mots de contenu.
90FFFD	Probabilité équivalant au 90e percentile de la distribution des fréquences des formes fléchies désambiguïsées.
90FFFDC	Probabilité équivalant au 90e percentile de la distribution des fréquences des formes fléchies désambiguïsées, uniquement pour les mots de contenu.
90FFFDG	Probabilité équivalant au 90e percentile de la distribution des fréquences des formes fléchies désambiguïsées, uniquement pour les mots grammaticaux.
90FFFDN	Probabilité équivalant au 90e percentile de la distribution des fréquences des formes fléchies désambiguïsées, uniquement pour les noms.
90FFFG	Probabilité équivalant au 90e percentile de la distribution des fréquences des formes fléchies, uniquement pour les mots grammaticaux.
90FFFN	Probabilité équivalant au 90e percentile de la distribution des fréquences des formes fléchies, uniquement pour les noms.
90FL	Probabilité équivalant au 90e percentile de la distribution des fréquences des lemmes.
90FLC	Probabilité équivalant au 90e percentile de la distribution des fréquences du lemme, uniquement pour les mots de contenu.
90FLG	Probabilité équivalant au 90e percentile de la distribution des fréquences du lemme, uniquement pour les mots grammaticaux.
90FLN	Probabilité équivalant au 90e percentile de la distribution des fréquences du lemme, uniquement pour les noms.
ADJ/ADV	Rapport des adjectifs aux adverbes.
ADJ/ALL	Proportion des adjectifs dans le texte.
ADJ/ALL_OTHERS	Rapport des adjectifs à l'ensemble des autres catégories de discours.
ADJ/ART	Rapport des adjectifs aux articles.
ADJ/CON	Rapport des adjectifs aux conjonctions.
ADJ/GRAM	Rapport des adjectifs aux mots grammaticaux.
ADJ/LEX	Rapport des adjectifs aux mots lexicaux.
ADJ/NAM	Rapport des adjectifs aux noms propres.
ADJ/NOM	Rapport des adjectifs aux noms communs.
ADJ/NOM+NAM	Rapport des adjectifs aux noms (communs et propres).
ADJ/PRE	Rapport des adjectifs aux prépositions.
ADJ/PRO	Rapport des adjectifs aux pronoms.
ADJ/VER	Rapport des adjectifs aux verbes.
ADV/ADJ	Rapport des adverbes aux adjectifs.
ADV/ALL	Proportion des adverbes dans le texte.
ADV/ALL_OTHERS	Rapport des adverbes à l'ensemble des autres catégories de discours.
ADV/ART	Rapport des adverbes aux articles.
ADV/CON	Rapport des adverbes aux conjonctions.

ADV/GRAM	Rapport des adverbes aux mots grammaticaux.
ADV/LEX	Rapport des adverbes aux mots lexicaux.
ADV/NAM	Rapport des adverbes aux noms propres.
ADV/NOM	Rapport des adverbes aux noms communs.
ADV/NOM+NAM	Rapport des adverbes aux noms (communs et propres).
ADV/PRE	Rapport des adverbes aux prépositions.
ADV/PRO	Rapport des adverbes aux pronoms.
ADV/VER	Rapport des adverbes aux verbes.
ANColl	Proportion des unités polylexicales nominales constituées d'un adjectif suivi d'un nom (*A N*).
ART/ADJ	Rapport des articles aux adjectifs.
ART/ADV	Rapport des articles aux adverbes.
ART/ALL	Proportion des articles dans le texte.
ART/ALL_OTHERS	Rapport des articles à l'ensemble des autres catégories de discours.
ART/CON	Rapport des articles aux conjonctions.
ART/GRAM	Rapport des articles aux mots grammaticaux.
ART/LEX	Rapport des articles aux mots lexicaux.
ART/NAM	Rapport des articles aux noms propres.
ART/NOM	Rapport des articles aux noms communs.
ART/NOM+NAM	Rapport des articles aux noms (communs et propres).
ART/PRE	Rapport des articles aux prépositions.
ART/PRO	Rapport des articles aux pronoms.
ART/VER	Rapport des articles aux verbes.
avLocalLsa_Lem	Cosinus moyen entre les phrases adjacentes du texte, calculé dans un espace obtenu par LSA, à partir des lemmes.
avLocalLsa_Tok	Cosinus moyen entre les phrases adjacentes du texte, calculé dans un espace obtenu par LSA, à partir des formes fléchies.
BINGUI	Présence de guillemets de dialogue.
CON/ADJ	Rapport des conjonctions aux adjectifs.
CON/ADV	Rapport des conjonctions aux adverbes.
CON/ALL	Proportion des conjonctions dans le texte.
CON/ALL_OTHERS	Rapport des conjonctions à l'ensemble des autres catégories de discours.
CON/ART	Rapport des conjonctions aux articles.
CON/GRAM	Rapport des conjonctions aux mots grammaticaux.
CON/LEX	Rapport des conjonctions aux mots lexicaux.
CON/NAM	Rapport des conjonctions aux noms propres.
CON/NOM	Rapport des conjonctions aux noms communs.
CON/NOM+NAM	Rapport des conjonctions aux noms (communs et propres).
CON/PRE	Rapport des conjonctions aux prépositions.
CON/PRO	Rapport des conjonctions aux pronoms.
CON/VER	Rapport des conjonctions aux verbes.

ConcDens	Nombre de propositions totales dans le texte divisé par le nombre de mots.
Cond	Présence ou absence du mode conditionnel dans le texte.
Cond_C	Proportion des verbes conjugués au conditionnel dans le texte.
coRefLocal1_Lem	Cosinus moyen entre les phrases adjacentes du texte, calculé dans un espace transformé par *tf-idf*, sur la base des lemmes.
coRefLocal1_Tok	Cosinus moyen entre les phrases adjacentes du texte, calculé dans un espace transformé par *tf-idf*, sur la base des formes fléchies.
freqCumNeigh_75	75^{e} percentile de la distribution des fréquences cumulées de tous les voisins de chaque mot.
freqCumNeigh_90	90^{e} percentile de la distribution des fréquences cumulées de tous les voisins de chaque mot.
Futur	Présence ou absence d'un temps du futur dans le texte.
Futur_C	Proportion des verbes conjugués au futur dans le texte.
gmeanNGProb-G	Moyenne géométrique des probabilités des bigrammes de Google.
gmeanNGProb-LS	Moyenne géométrique des probabilités des bigrammes du *Soir*.
GRAM/ADJ	Rapport des mots grammaticaux aux adjectifs.
GRAM/ADV	Rapport des mots grammaticaux aux adverbes.
GRAM/ALL	Proportion des mots grammaticaux dans le texte.
GRAM/ALL_OTHERS	Rapport des mots grammaticaux à l'ensemble des autres catégories de discours.
GRAM/ART	Rapport des mots grammaticaux aux articles.
GRAM/CON	Rapport des mots grammaticaux aux conjonctions.
GRAM/LEX	Rapport des mots grammaticaux aux mots lexicaux.
GRAM/NAM	Rapport des mots grammaticaux aux noms propres.
GRAM/NOM	Rapport des mots grammaticaux aux noms communs.
GRAM/NOM+NAM	Rapport des mots grammaticaux aux noms (communs et propres).
GRAM/PRE	Rapport des mots grammaticaux aux prépositions.
GRAM/PRO	Rapport des mots grammaticaux aux pronoms.
GRAM/VER	Rapport des mots grammaticaux aux verbes.
Imperatif	Présence ou absence d'un verbe à l'impératif dans le texte.
Imperatif_C	Proportion des verbes conjugués à l'impératif dans le texte.
Impf	Présence ou absence d'un verbe à l'imparfait dans le texte.
Impf_C	Proportion des verbes conjugués à l'imparfait dans le texte.
Infi	Présence ou absence d'un verbe à l'infinitif dans le texte.
Infi_C	Proportion des verbes à l'infinitif dans le texte.
IQFFF	Écart-interquartile des fréquences des formes fléchies.
IQFFFC	Écart-interquartile des fréquences des formes fléchies pour les mots de contenu.
IQFFFD	Écart-interquartile des fréquences des formes fléchies désambiguïsées.
IQFFFDC	Écart-interquartile des fréquences des formes fléchies désambiguïsée pour les mots de contenu.

IQFFFDG	Écart-interquartile des fréquences des formes fléchies désambiguïsée pour les mots grammaticaux.
IQFFFDN	Écart-interquartile des fréquences des formes fléchies désambiguïsée pour les noms.
IQFFFG	Écart-interquartile des fréquences des formes fléchies pour les mots grammaticaux.
IQFFFN	Écart-interquartile des fréquences des formes fléchies pour les noms.
IQFL	Écart-interquartile des fréquences des lemmes.
IQFLC	Écart-interquartile des fréquences du lemme pour les mots de contenu.
IQFLG	Écart-interquartile des fréquences du lemme pour les mots grammaticaux.
IQFLN	Écart-interquartile des fréquences du lemme pour les noms.
LenSentDaoust	Pourcentage de phrases de plus de 30 mots, repris à Daoust *et al.* (1996).
LenSentDaoust2	Pourcentage de phrases (définition alternative) de plus de 30 mots, repris à Daoust *et al.* (1996).
LenSentDolch	Mesure inspirée de Dolch (1948) et qui correspond à NWS_90 moins medNWS.
LenSentDolch2	Mesure inspirée de Dolch (1948) et qui correspond à NWS_902 moins medNWS2.
LEX/ADJ	Rapport des mots lexicaux aux adjectifs.
LEX/ADV	Rapport des mots lexicaux aux adverbes.
LEX/ALL	Proportion de mots lexicaux dans le texte.
LEX/ALL_OTHERS	Rapport des mots lexicaux à l'ensemble des autres catégories de discours.
LEX/ART	Rapport des mots lexicaux aux articles.
LEX/CON	Rapport des mots lexicaux aux conjonctions.
LEX/GRAM	Rapport des mots lexicaux aux mots grammaticaux.
LEX/NAM	Rapport des mots lexicaux aux noms propres.
LEX/NOM	Rapport des mots lexicaux aux noms communs.
LEX/NOM+NAM	Rapport des mots lexicaux aux noms (communs et propres).
LEX/PRE	Rapport des mots lexicaux aux prépositions.
LEX/PRO	Rapport des mots lexicaux aux pronoms.
LEX/VER	Rapport des mots lexicaux aux verbes.
mean_freqCumNeigh	Moyenne de la distribution de la fréquence cumulées des voisins de chaque mot.
mean_nbNeigh	Moyenne de la distribution du nombre de voisins par mots du texte.
mean_nbNeighMoreFreq	Moyenne de la distribution du nombre de voisins plus fréquents par mot.
mean_neighMoreFreq	Moyenne de la présence d'au moins un voisin plus fréquent par mot.
MeanFFF	Fréquence moyenne des formes fléchies.
MeanFFFA	Fréquence moyenne des formes fléchies pour les adjectifs.

MeanFFFC	Fréquence moyenne des formes fléchies pour les mots de contenu.
MeanFFFC-N	Fréquence moyenne des formes fléchies pour les mots de contenu, sauf les noms.
MeanFFFD	Fréquence moyenne des formes fléchies désambiguïsées.
MeanFFFDA	Fréquence moyenne des formes fléchies désambiguïsées pour les adjectifs.
MeanFFFDC	Fréquence moyenne des formes fléchies désambiguïsées pour les mots de contenu.
MeanFFFDC-N	Fréquence moyenne des formes fléchies désambiguïsées pour les mots de contenu, sauf les noms.
MeanFFFDG	Fréquence moyenne des formes fléchies désambiguïsées pour les mots grammaticaux.
MeanFFFDN	Fréquence moyenne des formes fléchies désambiguïsées pour les noms.
MeanFFFDV	Fréquence moyenne des formes fléchies désambiguïsées pour les verbes.
MeanFFFG	Fréquence moyenne des formes fléchies pour les mots grammaticaux.
MeanFFFN	Fréquence moyenne des formes fléchies pour les noms.
MeanFFFV	Fréquence moyenne des formes fléchies pour les verbes.
MeanFL	Fréquence moyenne des lemmes.
MeanFLA	Fréquence moyenne du lemme pour les adjectifs.
MeanFLC	Fréquence moyenne du lemme pour les mots de contenu.
MeanFLC-N	Fréquence moyenne du lemme pour les mots de contenu, sauf les noms.
MeanFLG	Fréquence moyenne du lemme pour les mots grammaticaux.
MeanFLN	Fréquence moyenne du lemme pour les noms.
MeanFLV	Fréquence moyenne du lemme pour les verbes.
meanNGProb-G	Moyenne des probabilités des bigrammes de Google.
meanNGProb-LS	Moyenne des probabilités des bigrammes du *Soir*.
meanNWS	Nombre moyen de mots par phrase.
MeanP2Coll	Probabilité moyenne des unités polylexicales nominales de taille 2.
MeanP3Coll	Probabilité moyenne des unités polylexicales nominales de taille 3.
MeanP4Coll	Probabilité moyenne des unités polylexicales nominales de taille 4.
MeanPColl	Probabilité moyenne des unités polylexicales nominales rencontrées dans le texte.
MeanSize	Taille moyenne (en nombre de mots) des unités polylexicales nominales présentes dans le texte.
med_freqCumNeigh	Médiane de la distribution de la fréquence cumulées des voisins de chaque mot.
med_nbNeigh	Médiane de la distribution du nombre de voisins par mots du texte.
med_nbNeighMoreFreq	Médiane de la distribution du nombre de voisins plus fréquents par mot.
MedianFFF	Médiane des fréquences des formes fléchies.

MedianFFFA	Médiane des fréquences des formes fléchies pour les adjectifs.
MedianFFFC	Médiane des fréquences des formes fléchies pour les mots de contenu.
MedianFFFC-N	Médiane des fréquences des formes fléchies pour les mots de contenu, sauf les noms.
MedianFFFD	Médiane des fréquences des formes fléchies désambiguïsées.
MedianFFFDA	Médiane des fréquences des formes fléchies désambiguïsées pour les adjectifs.
MedianFFFDC	Médiane des fréquences des formes fléchies désambiguïsées pour les mots de contenu.
MedianFFFDC-N	Médiane des fréquences des formes fléchies désambiguïsées pour les mots de contenu, sauf les noms.
MedianFFFDG	Médiane des fréquences des formes fléchies désambiguïsées pour les mots grammaticaux.
MedianFFFDN	Médiane des fréquences des formes fléchies désambiguïsées pour les noms.
MedianFFFDV	Médiane des fréquences des formes fléchies désambiguïsées pour les verbes.
MedianFFFG	Médiane des fréquences des formes fléchies pour les mots grammaticaux.
MedianFFFN	Médiane des fréquences des formes fléchies pour les noms.
MedianFFFV	Médiane des fréquences des formes fléchies pour les verbes.
MedianFL	Médiane des fréquences des lemmes.
MedianFLA	Médiane des fréquences du lemme pour les adjectifs.
MedianFLC	Médiane des fréquences du lemme pour les mots de contenu.
MedianFLC-N	Médiane des fréquences du lemme pour les mots de contenu, sauf les noms.
MedianFLG	Médiane des fréquences du lemme pour les mots grammaticaux.
MedianFLN	Médiane des fréquences du lemme pour les noms.
MedianFLV	Médiane des fréquences du lemme pour les verbes.
medianNGProb-G	Médiane des probabilités des bigrammes de Google.
medianNGProb-LS	Médiane des probabilités des bigrammes du *Soir*.
MedianP2Coll	Médianne de la distribution des probabilités des unités polylexicales nominales de taille 2.
MedianP3Coll	Médianne de la distribution des probabilités des unités polylexicales nominales de taille 3.
MedianP4Coll	Médianne de la distribution des probabilités des unités polylexicales nominales de taille 4.
MedianPColl	Médiane de la distribution des probabilités des unités polylexicales nominales rencontrées dans le texte.
medNWS	Nombre médian de mots par phrase.
medNWS2	Nombre médian de mots par phrase (définition alternative).
MedPCond-G	Médiane des probabilités conditionnelles du modèle bigramme basé sur Google.

MedPCond-LS	Médiane des probabilités conditionnelles du modèle bigramme basé sur *le Soir*.
ML1	Modèle unigramme basé sur les lemmes.
ML2	Modèle unigramme basé sur les formes fléchies.
ML3	Modèle unigramme basé sur les formes fléchies désambiguïsées.
MPCond-G	Moyenne des probabilités conditionnelles du modèle bigramme basé sur Google.
MPCond-LS	Moyenne des probabilités conditionnelles du modèle bigramme basé sur *le Soir*.
NAColl	Proportion d'unités polylexicales nominales de structure *N A*.
NAM/ADJ	Rapport des noms propres aux adjectifs.
NAM/ADV	Rapport des noms propres aux adverbes.
NAM/ALL	Proportion des noms propres dans le texte.
NAM/ALL_OTHERS	Rapport des noms propres à l'ensemble des autres catégories de discours.
NAM/ART	Rapport des noms propres aux articles.
NAM/CON	Rapport des noms propres aux conjonctions.
NAM/GRAM	Rapport des noms propres aux mots grammaticaux.
NAM/LEX	Rapport des noms propres aux mots lexicaux.
NAM/NOM	Rapport des noms propres aux noms communs.
NAM/NOM+NAM	Rapport des noms propres aux noms (communs et propres).
NAM/PRE	Rapport des noms propres aux prépositions.
NAM/PRO	Rapport des noms propres aux pronoms.
NAM/VER	Rapport des noms propres aux verbes.
nbNeigh_75	75^{e} percentile de la distribution du nombre de voisins par mots du texte.
nbNeigh_90	90^{e} percentile de la distribution du nombre de voisins par mots du texte.
nbNeighMoreFreq_75	75^{e} percentile de la distribution du nombre de voisins plus fréquents par mot.
nbNeighMoreFreq_90	90^{e} percentile de la distribution du nombre de voisins plus fréquents par mot.
NCPW	Proportion d'unités polylexicales nominales par rapport au nombre de mots.
NGprob75-G	75^{e} percentile de la distribution de probabilité des mots du texte, estimées à partir des bigrammes de Google.
NGprob75-LS	75^{e} percentile de la distribution de probabilité des mots du texte, estimées à partir des bigrammes du *Soir*.
NGprob90-G	90^{e} percentile de la distribution de probabilité des mots du texte, estimées à partir des bigrammes de Google.
NGprob90-LS	90^{e} percentile de la distribution de probabilité des mots du texte, estimées à partir des bigrammes du *Soir*.
NL3Q	Le nombre de lettres correspondant au troisième quartile (ou 75^{e} percentile) de la distribution de la longueur des mots du texte.

NL80P	Le nombre de lettres correspondant au 80^{e} percentile de la distribution de la longueur des mots du texte.
NL90P	Le nombre de lettres correspondant au 90^{e} percentile de la distribution de la longueur des mots du texte.
NLM	Le nombre moyen de lettres par mots.
NMLM	Le nombre médian de lettres par mots.
NMP	Nombre moyen de mots par phrase (définition alternative).
NNColl	Proportion d'unités polylexicales nominales dont la structure grammaticale correspond à la suite de catégories *N N*.
NOM/ADJ	Rapport des noms communs aux adjectifs.
NOM/ADV	Rapport des noms communs aux adverbes.
NOM/ALL	Proportion des noms communs dans le texte.
NOM/ALL_OTHERS	Rapport des noms communs à l'ensemble des autres catégories de discours.
NOM/ART	Rapport des noms communs aux articles.
NOM/CON	Rapport des noms communs aux conjonctions.
NOM/GRAM	Rapport des noms communs aux mots grammaticaux.
NOM/LEX	Rapport des noms communs aux mots lexicaux.
NOM/NAM	Rapport des noms communs aux noms propres.
NOM/NOM+NAM	Rapport des noms communs aux noms (communs et propres).
NOM/PRE	Rapport des noms communs aux prépositions.
NOM/PRO	Rapport des noms communs aux pronoms.
NOM/VER	Rapport des noms communs aux verbes.
NOM+NAM/ADJ	Rapport des noms aux adjectifs.
NOM+NAM/ADV	Rapport des noms aux adverbes.
NOM+NAM/ALL	Proportion des noms dans le texte.
NOM+NAM/ALL_OTHERS	Rapport des noms à l'ensemble des autres catégories de discours.
NOM+NAM/ART	Rapport des noms aux articles.
NOM+NAM/CON	Rapport des noms aux conjonctions.
NOM+NAM/GRAM	Rapport des noms aux mots grammaticaux.
NOM+NAM/LEX	Rapport des noms aux mots lexicaux.
NOM+NAM/NAM	Rapport des noms aux noms propres.
NOM+NAM/NOM	Rapport des noms aux noms communs.
NOM+NAM/PRE	Rapport des noms aux prépositions.
NOM+NAM/PRO	Rapport des noms aux pronoms.
NOM+NAM/VER	Rapport des noms aux verbes.
normTLProb-G	Vraisemblance du texte, calculée à partir du modèle bigramme basé sur les fréquences de Google.
normTLProb-LS	Vraisemblance du texte, calculée à partir du modèle bigramme basé sur les fréquences du *Soir*.
NormTTR-L	TTR normalisé calculé sur la base des formes fléchies.
NormTTR-W	TTR normalisé calculé à partir des lemmes.

NPNColl	Proportion d'unités polylexicales nominales dont la structure est *N PREP (DET) N*.
NWS_75	Longueur de la phrase correspondant au 75e percentile de la distribution du nombre de mots par phrase.
NWS_752	Longueur de la phrase correspondant au 75e percentile de la distribution du nombre de mots par phrase (définition alternative).
NWS_90	Longueur de la phrase correspondant au 90e percentile de la distribution du nombre de mots par phrase.
NWS_902	Longueur de la phrase correspondant au 90e percentile de la distribution du nombre de mots par phrase (définition alternative).
P2Coll75	Probabilité correspondant au 75e percentile de la distribution des probabilités des unités polylexicales nominales de taille 2.
P2Coll90	Probabilité correspondant au 90e percentile de la distribution des probabilités des unités polylexicales nominales de taille 2.
P3Coll75	Probabilité correspondant au 75e percentile de la distribution des probabilités des unités polylexicales nominales de taille 3.
P3Coll90	Probabilité correspondant au 90e percentile de la distribution des probabilités des unités polylexicales nominales de taille 3.
P4Coll75	Probabilité correspondant au 75e percentile de la distribution des probabilités des unités polylexicales nominales de taille 4.
P4Coll90	Probabilité correspondant au 90e percentile de la distribution des probabilités des unités polylexicales nominales de taille 4.
PA_Alterego1a	Pourcentage des mots lexicaux absents de la liste de 802 mots d'*Alter Ego 1*.
PA_Alterego1b	Pourcentage d'absents de la liste de 802 mots d'*Alter Ego 1* + 62 termes ajoutés.
PA_Alterego2a	Pourcentage des mots lexicaux absents des 1803 mots d'*Alter Ego* 1 et 2.
PA_Alterego2b	Pourcentage d'absents de la liste de 1803 mots d'*Alter Ego* 1 et 2 + 62 termes ajoutés.
PAGoug_1063	La proportion d'absents de la liste originale de 1 063 mots de Gougenheim.
PAGoug_2000	La proportion d'absents des 2 000 premiers mots de la liste étendue de Gougenheim.
PAGoug_3000	La proportion d'absents des 3 000 premiers mots de la liste étendue de Gougenheim.
PAGoug_4000	La proportion d'absents des 4 000 premiers mots de la liste étendue de Gougenheim.
PAGoug_5000	La proportion d'absents des 5 000 premiers mots de la liste étendue de Gougenheim.
PAGoug_6000	La proportion d'absents des 6 000 premiers mots de la liste étendue de Gougenheim.
PAGoug_7000	La proportion d'absents des 7 000 premiers mots de la liste étendue de Gougenheim.

PAGoug_8000	La proportion d'absents des 8 000 premiers mots de la liste étendue de Gougenheim.
PAGoug_8775	La proportion d'absents de la liste étendue de Gougenheim, qui comprend 8 775 mots.
PAGougU_1063	La proportion de formes différentes (« types ») absentes de la liste originale de 1 063 mots de Gougenheim.
PAGougU_2000	La proportion de formes différentes (« types » absentes des 2 000 premiers mots de la liste étendue de Gougenheim.
PAGougU_3000	La proportion de formes différentes (« types » absentes des 3 000 premiers mots de la liste étendue de Gougenheim.
PAGougU_4000	La proportion de formes différentes (« types » absentes des 4 000 premiers mots de la liste étendue de Gougenheim.
PAGougU_5000	La proportion de formes différentes (« types » absentes des 5 000 premiers mots de la liste étendue de Gougenheim.
PAGougU_6000	La proportion de formes différentes (« types » absentes des 6 000 premiers mots de la liste étendue de Gougenheim.
PAGougU_7000	La proportion de formes différentes (« types » absentes des 7 000 premiers mots de la liste étendue de Gougenheim.
PAGougU_8000	La proportion de formes différentes (« types » absentes des 8 000 premiers mots de la liste étendue de Gougenheim.
PAGougU_8775	La proportion de formes différentes (« types » absentes de la liste étendue de Gougenheim, qui comprend 8 775 mots.
PasseSim	Présence ou absence d'une forme conjuguée au passé simple dans le texte.
PasseSim_C	Proportion des verbes conjugués au passé simple dans le texte.
PAU_Alterego1a	Pourcentage des mots lexicaux différents (« types ») absents de la liste de 802 mots d'*Alter Ego 1*.
PAU_Alterego1b	Pourcentage de formes différentes (« types ») absentes de la liste de 802 mots d'*Alter Ego 1* + 62 termes ajoutés.
PAU_Alterego2a	Pourcentage des mots lexicaux différents (« types ») absents des 1803 mots d'*Alter Ego* 1 et 2.
PAU_Alterego2b	Pourcentage de formes différentes (« types ») absentes de la liste de 1803 mots d'*Alter Ego* 1 et 2 + 62 termes ajoutés.
PColl75	Probabilité correspondant au 75^{e} percentile de la distribution des probabilités des unités polylexicales nominales du texte.
Pcoll90	Probabilité correspondant au 90^{e} percentile de la distribution des probabilités des unités polylexicales nominales du texte.
PI	Pourcentage d'interjections dans le texte.
PM4	La proportion de mots de plus de 4 lettres.
PM5	La proportion de mots de plus de 5 lettres.
PM6	La proportion de mots de plus de 6 lettres.
PM7	La proportion de mots de plus de 7 lettres.
PM8	La proportion de mots de plus de 8 lettres.
PM9	La proportion de mots de plus de 9 lettres.

PM10	La proportion de mots de plus de 10 lettres.
PM11	La proportion de mots de plus de 11 lettres.
PM12	La proportion de mots de plus de 12 lettres.
PP1	Pourcentage de pronoms personnels de la 1re personne (je, me, moi, nous) sur le nombre de mots du texte.
PP1P2	Pourcentage de pron. pers. de la 1re et la 2e personne sur le nombre de mots du texte.
PP2	Pourcentage de pron. pers. de la 2e personne (tu, te, toi, vous) sur le nombre de mots du texte
PP2_P	Pourcentage de pron. pers. de la 2e personne au pluriel (vous) sur le nombre de mots du texte.
PP2_S	Pourcentage de pron. pers. de la 2e personne au singulier (tu, te, toi) sur le nombre de mots du texte.
PP3	Pourcentage de pron. pers. de la 3e personne (il, elle, le, la, lui, ils, elles, les, leur, eux) sur le nombre de mots du texte.
Ppasse	Présence ou absence d'une forme au participe passé dans le texte.
PPasse_C	Proportion des verbes conjugués au participe passé dans le texte.
PPD	Pourcentage de pronoms personnels de dialogue.
PPEI_1	Pourcentage de points d'exclamation et de points d'interrogation par rapport au nombre total de signes de ponctuation compris dans la liste suivante : (! ? . ;).
PPEI_2	Pourcentage de points d'exclamation et de points d'interrogation par rapport au nombre total de signes de ponctuation compris dans la liste suivante : (! ? .).
Ppres	Présence ou absence d'une forme au participe présent dans le texte.
PPres_C	Proportion des verbes conjugués au participe présent dans le texte.
PRE/ADJ	Rapport des prépositions aux adjectifs.
PRE/ADV	Rapport des prépositions aux adverbes.
PRE/ALL	Proportion des prépositions dans le texte.
PRE/ALL_OTHERS	Rapport des prépositions à l'ensemble des autres catégories de discours.
PRE/ART	Rapport des prépositions aux articles.
PRE/CON	Rapport des prépositions aux conjonctions.
PRE/GRAM	Rapport des prépositions aux mots grammaticaux.
PRE/LEX	Rapport des prépositions aux mots lexicaux.
PRE/NAM	Rapport des prépositions aux noms propres.
PRE/NOM	Rapport des prépositions aux noms communs.
PRE/NOM+NAM	Rapport des prépositions aux noms (communs + propres).
PRE/PRO	Rapport des prépositions aux pronoms.
PRE/VER	Rapport des prépositions aux verbes.
Pres	Présence ou absence d'une forme à l'indicatif présent dans le texte.
Pres_C	Proportion des verbes conjugués à l'indicatif présent dans le texte.
PRO/ADJ	Rapport des pronoms aux adjectifs.
PRO/ADV	Rapport des pronoms aux adverbes.

PRO/ALL	Proportion des pronoms dans le texte.
PRO/ALL_OTHERS	Rapport des pronoms à l'ensemble des autres catégories de discours.
PRO/ART	Rapport des pronoms aux articles.
PRO/CON	Rapport des pronoms aux conjonctions.
PRO/GRAM	Rapport des pronoms aux mots grammaticaux.
PRO/LEX	Rapport des pronoms aux mots lexicaux.
PRO/NAM	Rapport des pronoms aux noms propres.
PRO/NOM	Rapport des pronoms aux noms communs.
PRO/NOM+NAM	Rapport des pronoms aux noms (communs + propres).
PRO/PRE	Rapport des pronoms aux prépositions.
PRO/VER	Rapport des pronoms aux verbes.
PropP1	Proportion de pron. pers. de la 1re personne par rapport à l'ensemble des pron. pers.
PropP1P2	Proportion de pron. pers. de la 1re et la 2e personne par rapport à l'ensemble des pron. pers.
PropP2	Proportion de pron. pers. de la 2re personne par rapport à l'ensemble des pron. pers.
PropP2_P	Proportion de pron. pers. de la 2e personne au pluriel par rapport à l'ensemble des pron. pers.
PropP2_S	Proportion de pron. pers. de la 2e personne au singulier par rapport à l'ensemble des pron. pers.
PropP3	Proportion de pron. pers. de la 3re personne par rapport à l'ensemble des pron. pers.
ratioDaou	Ratio des verbes fléchis sur le nombre de mots du texte, inspiré de Daoust *et al.* (1996).
Subi	Présence ou absence d'une forme au subjonctif imparfait dans le texte.
Subi_C	Proportion des verbes conjugués au subjonctif imparfait dans le texte.
Subp	Présence ou absence d'une forme au subjonctif présent dans le texte.
Subp_C	Proportion des verbes conjugués au subjonctif présent dans le texte.
TTR-L	TTR calculé sur la base des formes fléchies.
TTR-W	TTR calculé à partir des lemmes.
VER/ADJ	Rapport des verbes aux adjectifs.
VER/ADV	Rapport des verbes aux adverbes.
VER/ALL	Porportion des verbes dans le texte.
VER/ALL_OTHERS	Rapport des verbes à l'ensemble des autres catégories de discours.
VER/ART	Rapport des verbes aux articles.
VER/CON	Rapport des verbes aux conjonctions.
VER/GRAM	Rapport des verbes aux mots grammaticaux.
VER/LEX	Rapport des verbes aux mots lexicaux.
VER/NAM	Rapport des verbes aux noms propres.
VER/NOM	Rapport des verbes aux noms communs.
VER/NOM+NAM	Rapport des verbes aux noms (communs + propres).

VER/PRE	Rapport des verbes aux prépositions.
VER/PRO	Rapport des verbes aux pronoms.

ANNEXE C

TABLEAU D'ANALYSE DES VARIABLES INDÉPENDANTES

	Test6CE				Test9CE			
Mesures de la fréquence lexicale								
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
MeanFL	-0.299^{3}	-0.282^{3}	$< 0,001$	0.037	-0.31^{3}	-0.304^{3}	$< 0,001$	0.076
MedianFL	-0.423^{3}	-0.389^{3}	$< 0,001$	0.014	-0.426^{3}	-0.401^{3}	$< 0,001$	0.299
IQFL	-0.006	0.02	$< 0,001$	0.583	-0.005	0.028	$< 0,001$	0.673
X75FL	-0.45^{3}	-0.579^{3}	$< 0,001$	$< 0,001$	-0.463^{3}	-0.572^{3}	$< 0,001$	$< 0,001$
X90FL	-0.445^{3}	-0.562^{3}	$< 0,001$	$< 0,001$	-0.468^{3}	-0.543^{3}	$< 0,001$	$< 0,001$
MeanFLC	-0.327^{3}	-0.297^{3}	$< 0,001$	0.128	-0.321^{3}	-0.297^{3}	$< 0,001$	0.41
MedianFLC	-0.259^{3}	-0.544^{3}	$< 0,001$	0.082	-0.255^{3}	-0.534^{3}	$< 0,001$	0.292
IQFLC	0.352^{3}	0.388^{3}	$< 0,001$	0.087	0.348^{3}	0.382^{3}	$< 0,001$	0.263
X75FLC	-0.32^{3}	-0.598^{3}	$< 0,001$	0.05	-0.399^{3}	-0.587^{3}	$< 0,001$	0.055
X90FLC	-0.343^{3}	-0.531^{3}	$< 0,001$	$< 0,001$	-0.402^{3}	-0.519^{3}	$< 0,001$	$< 0,001$
MeanFLG	-0.078	-0.039	$< 0,001$	0.32	-0.091^{1}	-0.052	$< 0,001$	0.397
MedianFLG	0.053	-0.06	$< 0,001$	0.502	0.032	-0.09^{1}	$< 0,001$	0.522
IQFLG	-0.013	-0.052	$< 0,001$	0.592	-0.058	-0.109^{2}	$< 0,001$	0.618
X75FLG	-0.234^{3}	-0.148^{2}	$< 0,001$	0.007	-0.256^{3}	-0.191^{3}	$< 0,001$	0.003
X90FLG	-0.244^{3}	-0.181^{2}	$< 0,001$	$< 0,001$	-0.299^{3}	-0.243^{3}	$< 0,001$	$< 0,001$
MeanFLN	-0.337^{3}	-0.35^{3}	$< 0,001$	0.093	-0.329^{3}	-0.33^{3}	$< 0,001$	0.413
MedianFLN	-0.326^{3}	-0.461^{3}	$< 0,001$	$< 0,001$	-0.337^{3}	-0.446^{3}	$< 0,001$	0.003
IQFLN	0.286^{3}	0.333^{3}	$< 0,001$	0.165	0.27^{3}	0.299^{3}	$< 0,001$	0.308
X75FLN	-0.318^{3}	-0.504^{3}	$< 0,001$	$< 0,001$	-0.312^{3}	-0.492^{3}	$< 0,001$	$< 0,001$
X90FLN	-0.255^{3}	-0.452^{3}	$< 0,001$	$< 0,001$	-0.248^{3}	-0.454^{3}	$< 0,001$	$< 0,001$
MeanFLV	-0.215^{3}	-0.171^{2}	$< 0,001$	0.03	-0.232^{3}	-0.189^{3}	$< 0,001$	0.198
MedianFLV	-0.362^{3}	-0.446^{3}	$< 0,001$	$< 0,001$	-0.373^{3}	-0.452^{3}	$< 0,001$	$< 0,001$
MeanFLA	-0.201^{3}	-0.133^{2}	$< 0,001$	0.453	-0.191^{3}	-0.119^{2}	$< 0,001$	0.644
MedianFLA	-0.258^{3}	-0.367^{3}	$< 0,001$	0.24	-0.24^{3}	-0.336^{3}	$< 0,001$	0.543
MeanFLC.N	-0.278^{3}	-0.256^{3}	$< 0,001$	0.1	-0.279^{3}	-0.257^{3}	$< 0,001$	0.175
MedianFLC.N	-0.338^{3}	-0.489^{3}	$< 0,001$	0.002	-0.337^{3}	-0.473^{3}	$< 0,001$	0.001
MeanFFF	-0.266^{3}	-0.238^{3}	$< 0,001$	0.013	-0.292^{3}	-0.277^{3}	$< 0,001$	$< 0,001$
MedianFFF	-0.363^{3}	-0.307^{3}	$< 0,001$	0.026	-0.377^{3}	-0.336^{3}	$< 0,001$	0.056
IQFFF	0.137^{1}	0.1	$< 0,001$	0.149	0.169^{3}	0.15^{3}	$< 0,001$	0.08
X75FFF	-0.372^{3}	-0.564^{3}	$< 0,001$	$< 0,001$	-0.411^{3}	-0.547^{3}	$< 0,001$	$< 0,001$
X90FFF	-0.363^{3}	-0.61^{3}	$< 0,001$	$< 0,001$	-0.322^{3}	-0.584^{3}	$< 0,001$	$< 0,001$
MeanFFFN	-0.254^{3}	-0.257^{3}	$< 0,001$	0.245	-0.22^{3}	-0.234^{3}	$< 0,001$	0.708
MedianFFFN	-0.359^{3}	-0.492^{3}	$< 0,001$	$< 0,001$	-0.361^{3}	-0.469^{3}	$< 0,001$	0.002
IQFFFN	0.321^{3}	0.342^{3}	$< 0,001$	0.193	0.284^{3}	0.307^{3}	$< 0,001$	0.413
X75FFFN	-0.282^{3}	-0.518^{3}	$< 0,001$	$< 0,001$	-0.286^{3}	-0.495^{3}	$< 0,001$	$< 0,001$
X90FFFN	-0.225^{3}	-0.528^{3}	$< 0,001$	0.004	-0.217^{3}	-0.518^{3}	$< 0,001$	$< 0,001$
MeanFFFC	-0.358^{3}	-0.319^{3}	$< 0,001$	$< 0,001$	-0.372^{3}	-0.341^{3}	$< 0,001$	0.004

MedianFFFC	-0.246^{3}	-0.56^{3}	$< 0,001$	0.099	-0.239^{3}	-0.56^{3}	$< 0,001$	0.083
IQFFFC	0.333^{3}	0.405^{3}	$< 0,001$	0.022	0.343^{3}	0.414^{3}	$< 0,001$	0.054
X75FFFC	-0.253^{3}	-0.624^{3}	$< 0,001$	0.047	-0.331^{3}	-0.619^{3}	$< 0,001$	0.077
X90FFFC	-0.319^{3}	-0.641^{3}	$< 0,001$	$< 0,001$	-0.246^{3}	-0.628^{3}	$< 0,001$	$< 0,001$
MeanFFFG	-0.052	-0.022	$< 0,001$	0.272	-0.073	-0.041	$< 0,001$	0.222
MedianFFFG	-0.19^{3}	-0.161^{3}	$< 0,001$	0.128	-0.2^{3}	-0.179^{3}	$< 0,001$	0.174
IQFFFG	-0.074	-0.08	$< 0,001$	0.505	-0.087^{1}	-0.093^{1}	$< 0,001$	0.573
X75FFFG	-0.212^{3}	-0.146^{2}	$< 0,001$	0.032	-0.238^{3}	-0.189^{3}	$< 0,001$	0.022
X90FFFG	-0.254^{3}	-0.191^{3}	$< 0,001$	0.03	-0.281^{3}	-0.226^{3}	$< 0,001$	0.079
MeanFFFV	-0.262^{3}	-0.179^{3}	$< 0,001$	$< 0,001$	-0.294^{3}	-0.217^{3}	$< 0,001$	$< 0,001$
MedianFFFV	-0.262^{3}	-0.179^{3}	$< 0,001$	$< 0,001$	-0.294^{3}	-0.217^{3}	$< 0,001$	$< 0,001$
MeanFFFA	-0.186^{3}	-0.088	$< 0,001$	0.303	-0.179^{3}	-0.086^{1}	$< 0,001$	0.594
MedianFFFA	-0.242^{3}	-0.442^{3}	$< 0,001$	0.065	-0.225^{3}	-0.425^{3}	$< 0,001$	0.559
MeanFFFC.N	-0.32^{3}	-0.278^{3}	$< 0,001$	$< 0,001$	-0.338^{3}	-0.304^{3}	$< 0,001$	$< 0,001$
MedianFFFC.N	-0.343^{3}	-0.509^{3}	$< 0,001$	0.005	-0.353^{3}	-0.514^{3}	$< 0,001$	$< 0,001$
MeanFFFD	-0.262^{3}	-0.235^{3}	$< 0,001$	0.006	-0.294^{3}	-0.275^{3}	$< 0,001$	$< 0,001$
MedianFFFD	-0.413^{3}	-0.375^{3}	$< 0,001$	$< 0,001$	-0.434^{3}	-0.4^{3}	$< 0,001$	$< 0,001$
IQFFFD	0.211^{3}	0.157^{2}	$< 0,001$	0.006	0.245^{3}	0.204^{3}	$< 0,001$	$< 0,001$
X75FFFD	-0.388^{3}	-0.596^{3}	$< 0,001$	$< 0,001$	-0.419^{3}	-0.593^{3}	$< 0,001$	$< 0,001$
X90FFFD	-0.422^{3}	-0.608^{3}	$< 0,001$	$< 0,001$	-0.439^{3}	-0.604^{3}	$< 0,001$	$< 0,001$
MeanFFFDC	-0.322^{3}	-0.281^{3}	$< 0,001$	$< 0,001$	-0.341^{3}	-0.307^{3}	$< 0,001$	0.003
MedianFFFDC	-0.219^{3}	-0.564^{3}	$< 0,001$	0.175	-0.207^{3}	-0.569^{3}	$< 0,001$	0.233
IQFFFDC	0.292^{3}	0.417^{3}	$< 0,001$	0.071	0.304^{3}	0.421^{3}	$< 0,001$	0.144
X75FFFDC	-0.296^{2}	-0.627^{3}	$< 0,001$	0.089	-0.367^{3}	-0.623^{3}	$< 0,001$	0.092
X90FFFDC	-0.404^{3}	-0.59^{3}	$< 0,001$	$< 0,001$	-0.381^{3}	-0.589^{3}	$< 0,001$	$< 0,001$
MeanFFFDG	-0.079	-0.05	$< 0,001$	0.178	-0.106^{1}	-0.071	$< 0,001$	0.081
MedianFFFDG	-0.227^{3}	-0.129^{2}	$< 0,001$	0.015	-0.236^{3}	-0.136^{3}	$< 0,001$	0.037
IQFFFDG	-0.098	-0.136^{2}	$< 0,001$	0.581	-0.115^{2}	-0.154^{3}	$< 0,001$	0.486
X75FFFDG	-0.256^{3}	-0.237^{3}	$< 0,001$	$< 0,001$	-0.297^{3}	-0.294^{3}	$< 0,001$	$< 0,001$
X90FFFDC.1	-0.316^{3}	-0.239^{3}	$< 0,001$	$< 0,001$	-0.353^{3}	-0.291^{3}	$< 0,001$	$< 0,001$
MeanFFFDN	-0.359^{3}	-0.371^{3}	$< 0,001$	0.133	-0.338^{3}	-0.338^{3}	$< 0,001$	0.487
MedianFFFDN	-0.354^{3}	-0.501^{3}	$< 0,001$	$< 0,001$	-0.363^{3}	-0.483^{3}	$< 0,001$	$< 0,001$
IQFFFDN	0.316^{3}	0.342^{3}	$< 0,001$	0.188	0.29^{3}	0.312^{3}	$< 0,001$	0.441
X75FFFDN	-0.288^{3}	-0.503^{3}	$< 0,001$	$< 0,001$	-0.283^{3}	-0.489^{3}	$< 0,001$	$< 0,001$
X90FFFDN	-0.228^{3}	-0.5^{3}	$< 0,001$	0.003	-0.221^{3}	-0.498^{3}	$< 0,001$	$< 0,001$
MeanFFFDV	-0.231^{3}	-0.128^{1}	$< 0,001$	$< 0,001$	-0.267^{3}	-0.168^{3}	$< 0,001$	$< 0,001$
MedianFFFDV	-0.251^{3}	-0.485^{3}	$< 0,001$	$< 0,001$	-0.268^{3}	-0.51^{3}	$< 0,001$	$< 0,001$
MeanFFFDA	-0.18^{3}	-0.083	$< 0,001$	0.308	-0.173^{3}	-0.081^{1}	$< 0,001$	0.563
MedianFFFDA	-0.234^{3}	-0.432^{3}	$< 0,001$	0.073	-0.219^{3}	-0.415^{3}	$< 0,001$	0.553
MeanFFFDC.N	-0.286^{3}	-0.238^{3}	$< 0,001$	$< 0,001$	-0.309^{3}	-0.268^{3}	$< 0,001$	$< 0,001$
MedianFFFDC.N	-0.315^{3}	-0.51^{3}	$< 0,001$	0.016	-0.321^{3}	-0.519^{3}	$< 0,001$	0.003
Proportion d'absents d'une liste de référence								
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
PAGoug_1063	0.578^{3}	0.577^{3}	$< 0,001$	0.008	0.562^{3}	0.574^{3}	$< 0,001$	0.235
PAGoug_2000	0.593^{3}	0.597^{3}	$< 0,001$	0.017	0.574^{3}	0.588^{3}	$< 0,001$	0.313
PAGoug_3000	0.584^{3}	0.591^{3}	0.015	0.018	0.567^{3}	0.586^{3}	$< 0,001$	0.317
PAGoug_4000	0.578^{3}	0.585^{3}	0.036	0.031	0.56^{3}	0.578^{3}	$< 0,001$	0.327
PAGoug_5000	0.567^{3}	0.578^{3}	0.029	0.132	0.543^{3}	0.565^{3}	$< 0,001$	0.203
PAGoug_6000	0.549^{3}	0.563^{3}	0.054	0.247	0.515^{3}	0.54^{3}	$< 0,001$	0.109
PAGoug_7000	0.525^{3}	0.543^{3}	0.024	0.267	0.486^{3}	0.517^{3}	$< 0,001$	0.077
PAGoug_8000	0.514^{3}	0.53^{3}	0.022	0.356	0.471^{3}	0.502^{3}	$< 0,001$	0.078
PAGoug_8775	0.502^{3}	0.519^{3}	0.016	0.374	0.451^{3}	0.486^{3}	$< 0,001$	0.032
PAGougU_1063	0.523^{3}	0.518^{3}	$< 0,001$	$< 0,001$	0.529^{3}	0.541^{3}	$< 0,001$	0.019
PAGougU_2000	0.542^{3}	0.547^{3}	0.002	$< 0,001$	0.546^{3}	0.564^{3}	$< 0,001$	0.035

PAGougU_3000	0.549^{3}	0.555^{3}	0.054	$< 0,001$	0.549^{3}	0.571^{3}	0.002	0.116
PAGougU_4000	0.544^{3}	0.556^{3}	0.089	0.001	0.548^{3}	0.572^{3}	$< 0,001$	0.193
PAGougU_5000	0.538^{3}	0.557^{3}	0.016	0.012	0.539^{3}	0.569^{3}	$< 0,001$	0.339
PAGougU_6000	0.522^{3}	0.544^{3}	0.007	0.112	0.515^{3}	0.548^{3}	$< 0,001$	0.448
PAGougU_7000	0.503^{3}	0.528^{3}	0.003	0.177	0.489^{3}	0.526^{3}	$< 0,001$	0.45
PAGougU_8000	0.491^{3}	0.518^{3}	0.006	0.253	0.475^{3}	0.514^{3}	$< 0,001$	0.494
PAGougU_8775	0.484^{3}	0.507^{3}	0.006	0.319	0.46^{3}	0.497^{3}	$< 0,001$	0.356
PA_Alterego1a	0.657^{3}	0.652^{3}	$< 0,001$	$< 0,001$	0.668^{3}	0.672^{3}	$< 0,001$	0.002
PAU_Alterego1a	0.093	0.105	$< 0,001$	0.002	0.129^{2}	0.142^{3}	$< 0,001$	0.004
PA_Alterego1b	0.649^{3}	0.636^{3}	$< 0,001$	$< 0,001$	0.659^{3}	0.656^{3}	$< 0,001$	0.001
PAU_Alterego1b	0.569^{3}	0.55^{3}	$< 0,001$	$< 0,001$	0.603^{3}	0.601^{3}	$< 0,001$	$< 0,001$
PA_Alterego2a	0.621^{3}	0.623^{3}	$< 0,001$	0.009	0.604^{3}	0.616^{3}	$< 0,001$	0.248
PAU_Alterego2a	0.03	0.057	$< 0,001$	0.028	0.045	0.072	$< 0,001$	0.016
PA_Alterego2b	0.631^{3}	0.627^{3}	$< 0,001$	0.005	0.618^{3}	0.623^{3}	$< 0,001$	0.198
PAU_Alterego2b	0.563^{3}	0.556^{3}	0.009	$< 0,001$	0.573^{3}	0.585^{3}	0.022	0.001
Modèles unigrammes								
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
ML1	-0.521^{3}	-0.522^{3}	0.031	0.076	-0.498^{3}	-0.513^{3}	0.038	0.538
ML2	-0.53^{3}	-0.52^{3}	$< 0,001$	0.013	-0.508^{3}	-0.507^{3}	0.022	0.229
ML3	-0.56^{3}	-0.546^{3}	$< 0,001$	$< 0,001$	-0.556^{3}	-0.552^{3}	$< 0,001$	0.026
Modèles n-grammes								
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
normTLProb-LS	-0.072	-0.022	$< 0,001$	0.177	-0.062	-0.011	$< 0,001$	0.566
normTLProb-G	-0.084	-0.011	$< 0,001$	0.31	-0.069	0.02	$< 0,001$	0.499
MPCond-LS	-0.196^{3}	-0.113^{1}	$< 0,001$	$< 0,001$	-0.222^{3}	-0.124^{2}	$< 0,001$	$< 0,001$
MPCond-G	-0.255^{3}	-0.186^{3}	$< 0,001$	$< 0,001$	-0.277^{3}	-0.198^{3}	$< 0,001$	$< 0,001$
MedPCond-LS	-0.08	-0.029	$< 0,001$	0.113	-0.049	0.003	$< 0,001$	0.138
MedPCond-G	-0.151^{3}	-0.056	$< 0,001$	0.386	-0.123^{3}	-0.027	$< 0,001$	0.42
meanNGProb-LS	0.28^{3}	0.303^{3}	$< 0,001$	0.162	-0.267^{3}	-0.13^{2}	$< 0,001$	0.342
meanNGProb-G	0.382^{3}	0.407^{3}	0.011	0.05	-0.244^{3}	-0.104^{1}	$< 0,001$	0.417
gmeanNGProb-LS	0.091	0.128^{2}	$< 0,001$	0.077	0.031	0.087^{1}	$< 0,001$	0.383
gmeanNGProb-G	-0.009	0.118^{1}	$< 0,001$	0.387	-0.014	0.127^{3}	$< 0,001$	0.676
medianNGProb-LS	-0.088	-0.079	$< 0,001$	0.42	-0.081^{1}	-0.065	$< 0,001$	0.676
medianNGProb-G	-0.002	0.114^{1}	$< 0,001$	0.39	0.011	0.135^{3}	$< 0,001$	0.433
NGprob75-LS	0.058	0.058	/	0.64	/	/	/	/
NGprob75-G	-0.096	-0.107^{1}	$< 0,001$	0.522	-0.081^{1}	-0.071	$< 0,001$	0.771
NGprob90-LS	0.058	0.058	/	0.64	/	/	/	/
NGprob90-G	-0.072	-0.065	$< 0,001$	0.523	-0.061	-0.051	$< 0,001$	0.674
Mesures de la diversité lexicale								
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
TTR_W	-0.283^{3}	-0.284^{3}	0.008	0.004	-0.251^{3}	-0.253^{3}	0.005	0.005
TTR_L	-0.269^{3}	-0.26^{3}	$< 0,001$	0.009	-0.25^{3}	-0.241^{3}	$< 0,001$	0.025
NormTTR_W	0.125^{2}	0.086	$< 0,001$	0.015	0.171^{3}	0.13^{3}	$< 0,001$	$< 0,001$
NormTTR_L	0.092	0.08	$< 0,001$	0.116	0.127^{2}	0.107^{2}	$< 0,001$	0.013
Le voisinage orthographique								
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
mean_nbNeigh	-0.086	-0.043	$< 0,001$	0.182	-0.096^{1}	-0.05	$< 0,001$	0.194
med_nbNeigh	-0.075	-0.07	$< 0,001$	0.358	-0.084^{1}	-0.072	$< 0,001$	0.353
nbNeigh_75	0.089	0.126^{1}	$< 0,001$	0.465	0.078	0.124^{2}	$< 0,001$	0.67
nbNeigh_90	-0.194^{3}	-0.099^{1}	$< 0,001$	0.099	-0.184^{3}	-0.112^{1}	$< 0,001$	0.271
mean_neighMoreFreq	-0.031	-0.037	$< 0,001$	0.429	-0.019	-0.031	$< 0,001$	0.178
mean_nbNeighMoreFreq	-0.165^{3}	-0.068	$< 0,001$	0.367	-0.143^{3}	-0.041	$< 0,001$	0.555
med_nbNeighMoreFreq	-0.219^{3}	-0.229^{3}	$< 0,001$	0.359	-0.175^{3}	-0.186^{3}	$< 0,001$	0.386
nbNeighMoreFreq_75	-0.104^{1}	-0.124^{2}	$< 0,001$	0.422	-0.075^{1}	-0.093^{1}	$< 0,001$	0.601

nbNeighMoreFreq_90	-0.13^{2}	-0.192^{3}	$< 0,001$	0.44	-0.108^{2}	-0.144^{3}	$< 0,001$	0.59
mean_freqCumNeigh	0.173^{3}	0.19^{3}	$< 0,001$	0.571	0.172^{3}	0.185^{3}	$< 0,001$	0.27
med_freqCumNeigh	-0.083	0.022	$< 0,001$	0.317	-0.105^{2}	-0.003	$< 0,001$	0.111
freqCumNeigh_75	-0.151^{1}	-0.306^{3}	$< 0,001$	0.21	-0.154^{3}	-0.292^{3}	$< 0,001$	0.164
freqCumNeigh_90	-0.145^{2}	-0.274^{3}	$< 0,001$	0.156	-0.158^{3}	-0.245^{3}	$< 0,001$	0.061
La longueur des mots								
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
NLM	0.479^{3}	0.483^{3}	0.028	0.084	0.431^{3}	0.44^{3}	0.048	0.027
NMLM	0.276^{3}	0.297^{3}	$< 0,001$	0.363	0.244^{3}	0.269^{3}	$< 0,001$	0.492
PM4	0.433^{3}	0.424^{3}	$< 0,001$	0.078	0.408^{3}	0.405^{3}	$< 0,001$	0.011
PM5	0.474^{3}	0.474^{3}	$< 0,001$	0.08	0.434^{3}	0.445^{3}	0.002	0.011
PM6	0.467^{3}	0.471^{3}	0.014	0.131	0.418^{3}	0.432^{3}	0.01	0.031
PM7	0.508^{3}	0.515^{3}	0.015	0.053	0.458^{3}	0.469^{3}	0.002	0.005
PM8	0.504^{3}	0.525^{3}	$< 0,001$	0.043	0.463^{3}	0.485^{3}	$< 0,001$	0.002
PM9	0.48^{3}	0.509^{3}	$< 0,001$	0.074	0.435^{3}	0.465^{3}	$< 0,001$	0.012
PM10	0.433^{3}	0.477^{3}	$< 0,001$	0.058	0.399^{3}	0.442^{3}	$< 0,001$	0.081
PM11	0.414^{3}	0.474^{3}	$< 0,001$	0.053	0.371^{3}	0.439^{3}	$< 0,001$	0.03
PM12	0.394^{3}	0.478^{3}	$< 0,001$	0.281	0.368^{3}	0.46^{3}	$< 0,001$	0.225
NL3Q	0.487^{3}	0.485^{3}	$< 0,001$	0.07	0.455^{3}	0.451^{3}	$< 0,001$	0.007
NL80P	0.491^{3}	0.489^{3}	$< 0,001$	0.075	0.453^{3}	0.452^{3}	$< 0,001$	0.026
NL90P	0.519^{3}	0.521^{3}	$< 0,001$	0.022	0.478^{3}	0.485^{3}	$< 0,001$	0.021
Mesures de la longueur des unités syntaxiques								
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
meanNWS	0.471^{3}	0.6^{3}	$< 0,001$	0.009	0.475^{3}	0.635^{3}	$< 0,001$	0.015
medNWS	0.413^{3}	0.539^{3}	$< 0,001$	0.022	0.417^{3}	0.57^{3}	$< 0,001$	0.03
NWS_75	0.435^{3}	0.577^{3}	$< 0,001$	0.01	0.443^{3}	0.62^{3}	$< 0,001$	0.049
NWS_90	0.444^{3}	0.588^{3}	$< 0,001$	0.004	0.459^{3}	0.633^{3}	$< 0,001$	0.026
LenSentDolch	0.339^{3}	0.462^{3}	$< 0,001$	0.015	0.366^{3}	0.501^{3}	$< 0,001$	0.119
LenSentDaoust	0.456^{3}	0.541^{3}	$< 0,001$	0.214	0.467^{3}	0.554^{3}	$< 0,001$	0.528
NMP	0.486^{3}	0.618^{3}	$< 0,001$	0.014	0.487^{3}	0.652^{3}	$< 0,001$	0.031
medNWS2	0.44^{3}	0.545^{3}	$< 0,001$	0.043	0.432^{3}	0.571^{3}	$< 0,001$	0.038
NWS_752	0.471^{3}	0.592^{3}	$< 0,001$	0.004	0.466^{3}	0.63^{3}	$< 0,001$	0.066
NWS_902	0.492^{3}	0.612^{3}	$< 0,001$	0.003	0.491^{3}	0.655^{3}	$< 0,001$	0.041
LenSentDolch2	0.384^{3}	0.489^{3}	$< 0,001$	0.008	0.401^{3}	0.533^{3}	$< 0,001$	0.237
LenSentDaoust2	0.469^{3}	0.56^{3}	$< 0,001$	0.324	0.477^{3}	0.572^{3}	$< 0,001$	0.6
Les ratios de catégories grammaticales								
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
ADJ.ADV	-0.141^{1}	0.026	$< 0,001$	0.466	-0.146^{3}	0.021	$< 0,001$	0.516
ADJ.ALL	0.026	0.102^{1}	$< 0,001$	0.146	0.013	0.085	$< 0,001$	0.135
ADJ.ALL_OTHERS	-0.004	0.102^{1}	$< 0,001$	0.139	-0.017	0.085	$< 0,001$	0.16
ADJ.ART	-0.149^{2}	-0.13^{2}	$< 0,001$	0.1	-0.141^{3}	-0.127^{2}	$< 0,001$	0.11
ADJ.CON	-0.118^{1}	0.012	$< 0,001$	0.378	-0.114^{2}	-0.038	$< 0,001$	0.566
ADJ.GRAM	-0.057	0.065	$< 0,001$	0.082	-0.071	0.039	$< 0,001$	0.149
ADJ.LEX	0.076	0.124^{1}	$< 0,001$	0.208	0.067	0.113^{1}	$< 0,001$	0.118
ADJ.NAM	0.012	0.171^{3}	$< 0,001$	0.52	0.034	0.211^{3}	$< 0,001$	0.586
ADJ.NOM	-0.11^{1}	-0.032	$< 0,001$	0.297	-0.115^{1}	-0.048	$< 0,001$	0.568
ADJ.NOM.NAM	-0.026	0.07	$< 0,001$	0.341	-0.021	0.067	$< 0,001$	0.112
ADJ.PRE	-0.123^{1}	-0.13^{2}	$< 0,001$	0.274	-0.143^{3}	-0.165^{3}	$< 0,001$	0.122
ADJ.PRO	-0.065^{1}	0.272^{3}	$< 0,001$	0.302	-0.06	0.259^{3}	$< 0,001$	0.649
ADJ.VER	-0.02	0.163^{3}	$< 0,001$	0.264	-0.024	0.141^{2}	$< 0,001$	0.457
ADV.ADJ	-0.091	-0.016	$< 0,001$	0.425	-0.088^{1}	-0.003	$< 0,001$	0.461
ADV.ALL	-0.006	0.034	$< 0,001$	0.51	-0.014	0.028	$< 0,001$	0.665
ADV.ALL_OTHERS	-0.022	0.034	$< 0,001$	0.503	-0.03	0.028	$< 0,001$	0.677
ADV.ART	-0.136^{2}	-0.084	$< 0,001$	0.161	-0.135^{3}	-0.075^{1}	$< 0,001$	0.127

ADV.CON	-0.07	0.017	$< 0,001$	0.426	-0.111^{2}	-0.022	$< 0,001$	0.245
ADV.GRAM	-0.057	0.011	$< 0,001$	0.4	-0.075	0.0	$< 0,001$	0.7
ADV.LEX	0.018	0.047	$< 0,001$	0.473	0.016	0.045	$< 0,001$	0.553
ADV.NAM	0.02	0.146^{2}	$< 0,001$	0.535	0.044	0.182^{3}	$< 0,001$	0.687
ADV.NOM	-0.132^{1}	-0.032	$< 0,001$	0.459	-0.117^{2}	-0.033	$< 0,001$	0.704
ADV.NOM.NAM	-0.061	0.026	$< 0,001$	0.348	-0.043	0.033	$< 0,001$	0.163
ADV.PRE	-0.144^{3}	-0.091	$< 0,001$	0.236	-0.156^{3}	-0.11^{2}	$< 0,001$	0.032
ADV.PRO	0.006	0.312^{3}	$< 0,001$	0.503	-0.009	0.284^{3}	$< 0,001$	0.704
ADV.VER	0.084	0.12	$< 0,001$	0.383	0.048	0.088^{1}	$< 0,001$	0.278
ART.ADJ	-0.025	0.138^{2}	$< 0,001$	0.523	-0.046	0.137^{2}	$< 0,001$	0.303
ART.ADV	-0.119^{2}	0.089	$< 0,001$	0.575	-0.11^{2}	0.087^{1}	$< 0,001$	0.7
ART.ALL	0.309^{3}	0.308^{3}	$< 0,001$	0.132	0.28^{3}	0.288^{3}	$< 0,001$	0.42
ART.ALL_OTHERS	0.29^{3}	0.308^{3}	0.003	0.143	0.261^{3}	0.288^{3}	0.031	0.475
ART.CON	-0.078	0.123^{2}	$< 0,001$	0.399	-0.078^{1}	0.067	$< 0,001$	0.478
ART.GRAM	0.273^{3}	0.279^{3}	$< 0,001$	0.123	0.234^{3}	0.247^{3}	$< 0,001$	0.436
ART.LEX	0.305^{3}	0.311^{3}	$< 0,001$	0.128	0.283^{3}	0.298^{3}	0.007	0.411
ART.NAM	0.047	0.224^{3}	$< 0,001$	0.564	0.069	0.262^{3}	$< 0,001$	0.642
ART.NOM	0.16	0.193^{3}	$< 0,001$	0.296	0.108	0.162^{3}	$< 0,001$	0.518
ART.NOM.NAM	0.306^{3}	0.29^{3}	$< 0,001$	0.062	0.295^{3}	0.279^{3}	$< 0,001$	0.099
ART.PRE	-0.087	0.018	$< 0,001$	0.411	-0.118^{2}	-0.027	$< 0,001$	0.214
ART.PRO	-0.02	0.326^{3}	$< 0,001$	0.319	-0.024	0.318^{3}	$< 0,001$	0.503
ART.VER	0.138	0.289^{3}	$< 0,001$	0.116	0.119^{2}	0.266^{3}	$< 0,001$	0.265
CON.ADJ	-0.06	-0.006	$< 0,001$	0.53	-0.066	0.05	$< 0,001$	0.589
CON.ADV	-0.112^{1}	0.021	$< 0,001$	0.597	-0.115^{2}	0.065	$< 0,001$	0.601
CON.ALL	0.074	0.073	$< 0,001$	0.293	0.127^{3}	0.132^{3}	$< 0,001$	0.068
CON.ALL_OTHERS	0.068	0.073	$< 0,001$	0.322	0.12^{3}	0.132^{3}	$< 0,001$	0.079
CON.ART	-0.134^{2}	-0.119^{1}	$< 0,001$	0.279	-0.126^{2}	-0.064	$< 0,001$	0.088
CON.GRAM	0.022	0.035	$< 0,001$	0.534	0.066	0.084^{1}	$< 0,001$	0.298
CON.LEX	0.087	0.092	$< 0,001$	0.17	0.143^{3}	0.156^{3}	$< 0,001$	0.036
CON.NAM	0.028	0.147^{2}	$< 0,001$	0.518	0.056	0.219^{3}	$< 0,001$	0.707
CON.NOM	-0.086	-0.031	$< 0,001$	0.546	-0.062	0.018	$< 0,001$	0.638
CON.NOM.NAM	0.009	0.048	$< 0,001$	0.139	0.065	0.117^{2}	$< 0,001$	0.003
CON.PRE	-0.164^{2}	-0.115^{1}	$< 0,001$	0.239	-0.163^{3}	-0.091^{2}	$< 0,001$	0.191
CON.PRO	-0.02	0.319^{3}	$< 0,001$	0.512	-0.011	0.335^{3}	$< 0,001$	0.578
CON.VER	0.072	0.173^{3}	$< 0,001$	0.558	0.07	0.2^{3}	$< 0,001$	0.183
GRAM.ADJ	-0.072	-0.065	$< 0,001$	0.509	-0.085^{1}	-0.039	$< 0,001$	0.277
GRAM.ADV	-0.15^{2}	-0.011	$< 0,001$	0.551	-0.151^{3}	0.0	$< 0,001$	0.563
GRAM.ALL	0.182^{2}	0.128^{1}	$< 0,001$	0.025	0.221^{3}	0.178^{3}	$< 0,001$	0.005
GRAM.ALL_OTHERS	0.152^{1}	0.128^{1}	$< 0,001$	0.047	0.191^{3}	0.178^{3}	$< 0,001$	0.023
GRAM.ART	-0.162^{3}	-0.279^{3}	$< 0,001$	0.137	-0.145^{3}	-0.247^{3}	$< 0,001$	0.062
GRAM.CON	-0.119^{1}	-0.035	$< 0,001$	0.371	-0.113^{2}	-0.084^{1}	$< 0,001$	0.338
GRAM.LEX	0.152^{1}	0.128^{1}	$< 0,001$	0.047	0.191^{3}	0.178^{3}	$< 0,001$	0.023
GRAM.NAM	0.031	0.155^{3}	$< 0,001$	0.531	0.056	0.213^{3}	$< 0,001$	0.634
GRAM.NOM	-0.167^{2}	-0.139^{2}	$< 0,001$	0.542	-0.135^{2}	-0.12^{2}	$< 0,001$	0.772
GRAM.NOM.NAM	-0.016	0.024	$< 0,001$	0.02	0.037	0.078	$< 0,001$	$< 0,001$
GRAM.PRE	-0.171^{3}	-0.304^{3}	$< 0,001$	0.269	-0.192^{3}	-0.316^{3}	$< 0,001$	0.023
GRAM.PRO	-0.028	0.342^{3}	$< 0,001$	0.344	-0.03	0.343^{3}	$< 0,001$	0.574
GRAM.VER	0.061	0.231^{3}	$< 0,001$	0.335	0.035	0.217^{3}	$< 0,001$	0.385
LEX.ADJ	-0.094	-0.124^{1}	$< 0,001$	0.394	-0.119^{2}	-0.113^{1}	$< 0,001$	0.119
LEX.ADV	-0.167^{3}	-0.047	$< 0,001$	0.513	-0.169^{3}	-0.045	$< 0,001$	0.491
LEX.ALL	-0.182^{2}	-0.128^{1}	$< 0,001$	0.025	-0.221^{3}	-0.178^{3}	$< 0,001$	0.005
LEX.ALL_OTHERS	-0.211^{3}	-0.128^{1}	$< 0,001$	0.011	-0.244^{3}	-0.178^{3}	$< 0,001$	$< 0,001$
LEX.ART	-0.158^{3}	-0.311^{3}	$< 0,001$	0.07	-0.158^{3}	-0.298^{3}	$< 0,001$	0.01
LEX.CON	-0.135^{2}	-0.092	$< 0,001$	0.333	-0.139^{3}	-0.156^{3}	$< 0,001$	0.222

LEX.GRAM	-0.211^{3}	-0.128^{1}	$< 0,001$	0.011	-0.244^{3}	-0.178^{3}	$< 0,001$	$< 0,001$
LEX.NAM	0.024	0.145^{2}	$< 0,001$	0.526	0.048	0.199^{3}	$< 0,001$	0.627
LEX.NOM	-0.231^{2}	-0.252^{3}	$< 0,001$	0.306	-0.193^{2}	-0.271^{3}	$< 0,001$	0.403
LEX.NOM.NAM	-0.126^{1}	-0.072	$< 0,001$	0.123	-0.085^{1}	-0.043	$< 0,001$	0.02
LEX.PRE	-0.158^{3}	-0.322^{3}	$< 0,001$	0.19	-0.183^{3}	-0.36^{3}	$< 0,001$	0.002
LEX.PRO	-0.059	0.286^{3}	$< 0,001$	0.33	-0.062	0.273^{3}	$< 0,001$	0.555
LEX.VER	-0.017	0.15^{2}	$< 0,001$	0.199	-0.031	0.106^{1}	$< 0,001$	0.406
NAM.ADJ	-0.134^{2}	-0.172^{3}	$< 0,001$	0.25	-0.14^{3}	-0.211^{3}	$< 0,001$	0.025
NAM.ADV	-0.183^{3}	-0.125^{2}	$< 0,001$	0.159	-0.183^{3}	-0.156^{3}	$< 0,001$	0.021
NAM.ALL	-0.23^{3}	-0.147^{2}	$< 0,001$	0.007	-0.282^{3}	-0.202^{3}	$< 0,001$	$< 0,001$
NAM.ALL_OTHERS	-0.232^{3}	-0.147^{2}	$< 0,001$	0.003	-0.281^{3}	-0.202^{3}	$< 0,001$	$< 0,001$
NAM.ART	-0.129^{2}	-0.22^{3}	$< 0,001$	0.087	-0.139^{3}	-0.259^{3}	$< 0,001$	0.003
NAM.CON	-0.162^{3}	-0.135^{2}	$< 0,001$	0.083	-0.183^{3}	-0.199^{3}	$< 0,001$	0.012
NAM.GRAM	-0.232^{3}	-0.152^{3}	$< 0,001$	0.002	-0.278^{3}	-0.21^{3}	$< 0,001$	$< 0,001$
NAM.LEX	-0.222^{3}	-0.143^{2}	$< 0,001$	0.012	-0.274^{3}	-0.196^{3}	$< 0,001$	$< 0,001$
NAM.NOM	-0.203^{2}	-0.18^{3}	$< 0,001$	0.06	-0.185^{2}	-0.23^{3}	$< 0,001$	0.1
NAM.NOM.NAM	-0.253^{3}	-0.18^{3}	$< 0,001$	0.02	-0.296^{3}	-0.23^{3}	$< 0,001$	$< 0,001$
NAM.PRE	-0.137^{2}	-0.233^{3}	$< 0,001$	0.099	-0.136^{3}	-0.29^{3}	$< 0,001$	0.009
NAM.PRO	-0.086	0.034	$< 0,001$	0.389	-0.097^{2}	-0.018	$< 0,001$	0.286
NAM.VER	-0.153^{3}	-0.089	$< 0,001$	0.023	-0.195^{3}	-0.15^{3}	$< 0,001$	$< 0,001$
NOM.NAM.ADJ	-0.09	-0.07	$< 0,001$	0.371	-0.12^{2}	-0.067	$< 0,001$	0.057
NOM.NAM.ADV	-0.164^{3}	-0.026	$< 0,001$	0.501	-0.161^{3}	-0.033	$< 0,001$	0.509
NOM.NAM.ALL	-0.006	0.03	$< 0,001$	0.029	-0.056	-0.015	$< 0,001$	$< 0,001$
NOM.NAM.ALL_OTHERS	-0.048	0.03	$< 0,001$	0.02	-0.095^{1}	-0.015	$< 0,001$	$< 0,001$
NOM.NAM.ART	-0.143^{2}	-0.29^{3}	$< 0,001$	0.067	-0.148^{3}	-0.279^{3}	$< 0,001$	0.004
NOM.NAM.CON	-0.136^{2}	-0.047	$< 0,001$	0.261	-0.131^{3}	-0.117^{2}	$< 0,001$	0.178
NOM.NAM.GRAM	-0.118^{1}	-0.024	$< 0,001$	0.001	-0.164^{3}	-0.078	$< 0,001$	$< 0,001$
NOM.NAM.LEX	0.068	0.072	0.009	0.158	0.029	0.043	0.002	0.008
NOM.NAM.NAM	0.035	0.184^{3}	$< 0,001$	0.546	0.058	0.233^{3}	$< 0,001$	0.634
NOM.NAM.NOM	-0.203^{2}	-0.18^{3}	$< 0,001$	0.059	-0.185^{2}	-0.23^{3}	$< 0,001$	0.098
NOM.NAM.PRE	-0.144^{2}	-0.289^{3}	$< 0,001$	0.181	-0.166^{3}	-0.345^{3}	$< 0,001$	$< 0,001$
NOM.NAM.PRO	-0.065^{1}	0.259^{3}	$< 0,001$	0.332	-0.071^{1}	0.238^{3}	$< 0,001$	0.482
NOM.NAM.VER	-0.027	0.135^{2}	$< 0,001$	0.218	-0.039	0.092^{1}	$< 0,001$	0.461
NOM.ADJ	-0.064	0.032	$< 0,001$	0.446	-0.091^{1}	0.048	$< 0,001$	0.136
NOM.ADV	-0.141^{2}	0.032	$< 0,001$	0.579	-0.134^{3}	0.034	$< 0,001$	0.592
NOM.ALL	0.213^{3}	0.219^{3}	$< 0,001$	0.435	0.212^{3}	0.221^{3}	$< 0,001$	0.595
NOM.ALL_OTHERS	0.18^{2}	0.219^{3}	$< 0,001$	0.434	0.183^{3}	0.221^{3}	$< 0,001$	0.58
NOM.ART	-0.157^{2}	-0.193^{3}	$< 0,001$	0.077	-0.138^{3}	-0.162^{3}	$< 0,001$	0.053
NOM.CON	-0.11^{1}	0.032	$< 0,001$	0.357	-0.101^{2}	-0.017	$< 0,001$	0.335
NOM.GRAM	0.067	0.139^{2}	$< 0,001$	0.369	0.047	0.12^{2}	$< 0,001$	0.538
NOM.LEX	0.256^{3}	0.252^{3}	$< 0,001$	0.266	0.264^{3}	0.271^{3}	$< 0,001$	0.53
NOM.NAM	0.035	0.183^{3}	$< 0,001$	0.546	0.057	0.233^{3}	$< 0,001$	0.634
NOM.NOM.NAM	0.253^{3}	0.18^{3}	$< 0,001$	0.02	0.296^{3}	0.23^{3}	$< 0,001$	$< 0,001$
NOM.PRE	-0.146^{2}	-0.16^{2}	$< 0,001$	0.269	-0.159^{3}	-0.203^{3}	$< 0,001$	0.015
NOM.PRO	-0.05	0.305^{3}	$< 0,001$	0.371	-0.054	0.297^{3}	$< 0,001$	0.56
NOM.VER	0.046	0.226^{3}	$< 0,001$	0.421	0.016	0.209^{3}	$< 0,001$	0.607
PRE.ADJ	-0.033	0.139^{2}	$< 0,001$	0.573	-0.039	0.176^{3}	$< 0,001$	0.484
PRE.ADV	-0.128^{1}	0.097	$< 0,001$	0.525	-0.119^{3}	0.126^{3}	$< 0,001$	0.686
PRE.ALL	0.341^{3}	0.324^{3}	$< 0,001$	0.192	0.364^{3}	0.353^{3}	$< 0,001$	0.009
PRE.ALL_OTHERS	0.323^{3}	0.324^{3}	$< 0,001$	0.207	0.347^{3}	0.353^{3}	$< 0,001$	0.02
PRE.ART	-0.089	-0.005	$< 0,001$	0.439	-0.077	0.046	$< 0,001$	0.464
PRE.CON	-0.08	0.119^{1}	$< 0,001$	0.472	-0.067	0.105^{2}	$< 0,001$	0.634
PRE.GRAM	0.282^{3}	0.304^{3}	$< 0,001$	0.324	0.301^{3}	0.316^{3}	$< 0,001$	0.035
PRE.LEX	0.342^{3}	0.322^{3}	$< 0,001$	0.16	0.367^{3}	0.36^{3}	$< 0,001$	0.009

PRE.NAM	0.052	0.241^{3}	$< 0,001$	0.554	0.076^{1}	0.306^{3}	$< 0,001$	0.649
PRE.NOM	0.093	0.16^{2}	$< 0,001$	0.423	0.074	0.203^{3}	$< 0,001$	0.616
PRE.NOM.NAM	0.301^{3}	0.289^{3}	$< 0,001$	0.009	0.354^{3}	0.345^{3}	$< 0,001$	$< 0,001$
PRE.PRO	-0.034	0.345^{3}	$< 0,001$	0.35	-0.038	0.349^{3}	$< 0,001$	0.597
PRE.VER	0.099	0.313^{3}	$< 0,001$	0.507	0.065	0.316^{3}	$< 0,001$	0.508
PRO.ADJ	-0.114^{1}	-0.272^{3}	$< 0,001$	0.45	-0.123^{2}	-0.257^{3}	$< 0,001$	0.221
PRO.ADV	-0.173^{3}	-0.298^{3}	$< 0,001$	0.398	-0.189^{3}	-0.274^{3}	$< 0,001$	0.1
PRO.ALL	-0.347^{3}	-0.313^{3}	$< 0,001$	0.136	-0.328^{3}	-0.305^{3}	$< 0,001$	0.161
PRO.ALL_OTHERS	-0.355^{3}	-0.313^{3}	$< 0,001$	0.148	-0.335^{3}	-0.305^{3}	$< 0,001$	0.139
PRO.ART	-0.165^{3}	-0.326^{3}	$< 0,001$	0.136	-0.153^{3}	-0.317^{3}	$< 0,001$	0.05
PRO.CON	-0.171^{3}	-0.307^{3}	$< 0,001$	0.258	-0.189^{3}	-0.323^{3}	$< 0,001$	0.028
PRO.GRAM	-0.389^{3}	-0.342^{3}	$< 0,001$	0.089	-0.382^{3}	-0.343^{3}	$< 0,001$	0.012
PRO.LEX	-0.316^{3}	-0.286^{3}	$< 0,001$	0.17	-0.29^{3}	-0.273^{3}	$< 0,001$	0.199
PRO.NAM	-0.025	-0.025	$< 0,001$	0.444	0.003	0.024	$< 0,001$	0.671
PRO.NOM	-0.276^{3}	-0.305^{3}	$< 0,001$	0.284	-0.223^{3}	-0.297^{3}	$< 0,001$	0.305
PRO.NOM.NAM	-0.28^{3}	-0.259^{3}	$< 0,001$	0.157	-0.24^{3}	-0.238^{3}	$< 0,001$	0.156
PRO.PRE	-0.181^{3}	-0.345^{3}	$< 0,001$	0.226	-0.194^{3}	-0.349^{3}	$< 0,001$	0.021
PRO.VER	-0.298^{3}	-0.277^{3}	$< 0,001$	0.095	-0.31^{3}	-0.294^{3}	$< 0,001$	0.062
VER.ADJ	-0.094^{1}	-0.163^{3}	$< 0,001$	0.448	-0.106^{2}	-0.141^{2}	$< 0,001$	0.241
VER.ADV	-0.165^{3}	-0.117	$< 0,001$	0.534	-0.172^{3}	-0.085^{1}	$< 0,001$	0.318
VER.ALL	-0.199^{3}	-0.193^{3}	$< 0,001$	0.066	-0.159^{3}	-0.16^{3}	$< 0,001$	0.067
VER.ALL_OTHERS	-0.214^{3}	-0.193^{3}	$< 0,001$	0.069	-0.172^{3}	-0.16^{3}	$< 0,001$	0.075
VER.ART	-0.172^{3}	-0.289^{3}	$< 0,001$	0.179	-0.16^{3}	-0.266^{3}	$< 0,001$	0.033
VER.CON	-0.133^{2}	-0.172^{3}	$< 0,001$	0.451	-0.143^{3}	-0.199^{3}	$< 0,001$	0.256
VER.GRAM	-0.25^{3}	-0.231^{3}	$< 0,001$	0.384	-0.234^{3}	-0.217^{3}	$< 0,001$	0.321
VER.LEX	-0.156^{2}	-0.15^{2}	$< 0,001$	0.017	-0.104^{1}	-0.106^{1}	$< 0,001$	0.006
VER.NAM	0.012	0.094^{1}	$< 0,001$	0.463	0.039	0.155^{3}	$< 0,001$	0.664
VER.NOM	-0.219^{2}	-0.226^{3}	$< 0,001$	0.364	-0.173^{2}	-0.209^{3}	$< 0,001$	0.71
VER.NOM.NAM	-0.162^{2}	-0.135^{2}	$< 0,001$	0.028	-0.111^{2}	-0.092^{1}	$< 0,001$	0.013
VER.PRE	-0.195^{3}	-0.313^{3}	$< 0,001$	0.268	-0.19^{3}	-0.316^{3}	$< 0,001$	0.021
Les aspects verbaux								
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
Cond	0.227^{3}	0.227^{3}	$< 0,001$	$< 0,001$	0.25^{3}	0.249^{3}	$< 0,001$	0.02
Futur	0.253^{3}	0.252^{3}	$< 0,001$	$< 0,001$	0.272^{3}	0.271^{3}	$< 0,001$	0.002
Imperatif	0.019	0.019	$< 0,001$	0.391	0.022	0.022	$< 0,001$	0.671
Impf	0.272^{3}	0.272^{3}	$< 0,001$	$< 0,001$	0.317^{3}	0.317^{3}	$< 0,001$	$< 0,001$
Infi	0.342^{3}	0.341^{3}	$< 0,001$	$< 0,001$	0.375^{3}	0.374^{3}	$< 0,001$	$< 0,001$
PPasse	0.388^{3}	0.388^{3}	$< 0,001$	$< 0,001$	0.425^{3}	0.424^{3}	$< 0,001$	$< 0,001$
PPres	0.44^{3}	0.44^{3}	$< 0,001$	0.003	0.463^{3}	0.463^{3}	$< 0,001$	0.023
Pres	0.078	0.078	$< 0,001$	0.406	0.059	0.059	$< 0,001$	0.69
PasseSim	0.146^{2}	0.146^{2}	$< 0,001$	0.017	0.155^{3}	0.155^{3}	$< 0,001$	0.125
Subi	0.049	0.049	$< 0,001$	0.608	0.076^{1}	0.076^{1}	$< 0,001$	0.575
Subp	0.266^{3}	0.266^{3}	$< 0,001$	0.423	0.247^{3}	0.247^{3}	$< 0,001$	0.438
Cond_C	0.023	0.192^{3}	$< 0,001$	0.197	0.043	0.218^{3}	$< 0,001$	0.252
Futur_C	0.065	0.219^{3}	$< 0,001$	0.086	0.064	0.239^{3}	$< 0,001$	0.021
Imperatif_C	-0.022	0.019	$< 0,001$	0.413	-0.032	0.021	$< 0,001$	0.681
Impf_C	0.024	0.223^{3}	$< 0,001$	0.076	0.088^{1}	0.275^{3}	$< 0,001$	$< 0,001$
Infi_C	0.232^{3}	0.282^{3}	$< 0,001$	0.069	0.273^{3}	0.331^{3}	$< 0,001$	0.006
PPasse_C	0.262^{3}	0.304^{3}	$< 0,001$	0.003	0.31^{3}	0.366^{3}	$< 0,001$	$< 0,001$
PPres_C	0.252^{3}	0.41^{3}	$< 0,001$	0.123	0.246^{3}	0.432^{3}	$< 0,001$	0.205
Pres_C	-0.355^{3}	-0.337^{3}	$< 0,001$	$< 0,001$	-0.439^{3}	-0.433^{3}	$< 0,001$	$< 0,001$
PasseSim_C	-0.012	0.121^{1}	$< 0,001$	0.086	-0.002	0.13^{3}	$< 0,001$	0.039
Subi_C	-0.023	0.047	$< 0,001$	0.494	0.007	0.074^{1}	$< 0,001$	0.409
Subp_C	-0.07	0.091	$< 0,001$	0.263	-0.073^{1}	0.09^{1}	$< 0,001$	0.301

ratioDaou	-0.288^{3}	-0.287^{3}	$< 0,001$	0.193	-0.258^{3}	-0.264^{3}	$< 0,001$	0.22
Taux d'abstraction et de personnalisation du texte								
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
PP1	-0.329^{3}	-0.224^{3}	$< 0,001$	0.044	-0.324^{3}	-0.235^{3}	$< 0,001$	0.012
PropP1	-0.153^{3}	-0.135^{2}	$< 0,001$	0.096	-0.162^{3}	-0.148^{3}	$< 0,001$	0.117
PP2	-0.338^{3}	-0.325^{3}	$< 0,001$	0.095	-0.333^{3}	-0.325^{3}	$< 0,001$	0.058
PropP2	-0.306^{3}	-0.306^{3}	$< 0,001$	0.347	-0.299^{3}	-0.303^{3}	$< 0,001$	0.251
PP3	-0.179^{3}	-0.063	$< 0,001$	0.411	-0.165^{3}	-0.042	$< 0,001$	0.57
PropP3	0.315^{3}	0.302^{3}	$< 0,001$	0.319	0.308^{3}	0.297^{3}	$< 0,001$	0.147
PP2_S	-0.229^{3}	-0.21^{3}	$< 0,001$	0.138	-0.22^{3}	-0.213^{3}	$< 0,001$	0.121
PropP2_S	-0.221^{3}	-0.209^{3}	$< 0,001$	0.21	-0.208^{3}	-0.212^{3}	$< 0,001$	0.138
PP2_P	-0.249^{3}	-0.246^{3}	$< 0,001$	0.233	-0.248^{3}	-0.242^{3}	$< 0,001$	0.17
PropP2_P	-0.223^{3}	-0.232^{3}	$< 0,001$	0.311	-0.217^{3}	-0.225^{3}	$< 0,001$	0.139
PP1P2	-0.408^{3}	-0.333^{3}	$< 0,001$	0.008	-0.405^{3}	-0.346^{3}	$< 0,001$	$< 0,001$
PP1P2.1	-0.408^{3}	-0.333^{3}	$< 0,001$	0.008	-0.405^{3}	-0.346^{3}	$< 0,001$	$< 0,001$
PropP1P2	-0.305^{3}	-0.282^{3}	$< 0,001$	0.069	-0.315^{3}	-0.298^{3}	$< 0,001$	0.009
Densité des idées								
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
ConcDens	0.25^{3}	0.253^{3}	$< 0,001$	0.093	0.282^{3}	0.277^{3}	$< 0,001$	$< 0,001$
Le taux de cohérence								
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
coRefLocal1_Tok	-0.183^{3}	-0.054	$< 0,001$	0.178	-0.167^{3}	-0.057	$< 0,001$	0.14
coRefLocal1_Lem	-0.116^{1}	0.001	$< 0,001$	0.09	-0.124^{3}	-0.005	$< 0,001$	0.066
avLocalLsa_Tok	0.662^{3}	0.682^{3}	$< 0,001$	$< 0,001$	0.59^{3}	0.603^{3}	$< 0,001$	$< 0,001$
avLocalLsa_Lem	0.657^{3}	0.671^{3}	$< 0,001$	$< 0,001$	0.592^{3}	0.599^{3}	$< 0,001$	$< 0,001$
Le type de texte								
	r	ρ	$W(p)$	$F(p)$	r	ρ	$W(p)$	$F(p)$
PPD	-0.396^{3}	-0.318^{3}	$< 0,001$	0.026	-0.401^{3}	-0.337^{3}	$< 0,001$	$< 0,001$
PI	-0.284^{3}	-0.245^{3}	$< 0,001$	$< 0,001$	-0.309^{3}	-0.269^{3}	$< 0,001$	$< 0,001$
PPEI_1	-0.212^{3}	-0.103	$< 0,001$	$< 0,001$	-0.253^{3}	-0.166^{3}	$< 0,001$	0.002
PPEI_2	-0.261^{3}	-0.139^{1}	$< 0,001$	0.001	-0.291^{3}	-0.199^{3}	$< 0,001$	$< 0,001$
BINGUI	0.462^{3}	0.462^{3}	$< 0,001$	0.018	0.45^{3}	0.45^{3}	$< 0,001$	0.311

	Le Soir				Google			
Seuils	0	19	25	43	0	139	4000	9931
NCPW	0.242^{3}	0.183^{3}	0.146^{2}	0.135^{2}	0.199^{3}	0.16^{3}	0.178^{3}	0.199^{3}
MeanSize	0.055	0.086	0.066	0.061	-0.017	-0.016	0.008	-0.013
NNColl	0.053	0.167^{3}	0.18^{3}	0.157^{3}	0.052	0.073	0.053	0.045
NPNColl	0.053	0.115^{1}	0.125^{2}	0.173^{3}	0.037	0.163^{3}	0.21^{3}	0.214^{3}
ANColl	0.006	0.069	0.079	0.071	-0.003	0.109^{1}	0.141^{2}	0.142^{3}
NAColl	0.283^{3}	0.282^{3}	0.271^{3}	0.286^{3}	0.31^{3}	0.325^{3}	0.262^{3}	0.253^{3}
MeanPColl	0.116^{2}	0.077	-0.066	0.05	0.254^{3}	0.18^{3}	0.109^{1}	0.098
MedianPColl	-0.083	-0.108^{1}	-0.075	-0.044	0.032	0.047	0.047	0.062
PColl75	-0.225^{3}	-0.261^{3}	-0.202^{3}	-0.15^{1}	-0.131^{2}	-0.064	-0.022	0.008
PColl90	-0.375^{3}	-0.367^{3}	-0.296^{3}	-0.207^{3}	-0.341^{3}	-0.17^{2}	-0.05	-0.006
MeanP2Coll	0.078	0.028	0.001	-0.028	0.203^{3}	0.09	0.054	0.037
MedianP2Coll	-0.103^{1}	-0.069	-0.096	-0.073	-0.088^{1}	-0.034	0.018	0.044
P2Coll75	-0.236^{3}	-0.172^{3}	-0.188^{3}	-0.15^{1}	-0.245^{3}	-0.114^{1}	-0.029	0.022
P2Coll90	-0.284^{3}	-0.241^{3}	-0.251^{3}	-0.21^{3}	-0.283^{3}	-0.162^{2}	-0.055	0.005
MeanP3Coll	0.093	0.0	-0.046	-0.018	0.168^{3}	0.1	0.1	0.057
MedianP3Coll	-0.167^{3}	-0.086	-0.101	-0.064	-0.083	0.075	0.094	0.059
P3Coll75	-0.235^{3}	-0.163^{2}	-0.158^{2}	-0.106	-0.174^{3}	0.045	0.081	0.045
P3Coll90	-0.242^{3}	-0.183^{3}	-0.192^{3}	-0.131^{1}	-0.183^{3}	0.025	0.069	0.039
MeanP4Coll	0.12^{1}	0.044	-0.051	-0.146	0.167^{2}	0.084	0.035	0.09
MedianP4Coll	-0.098	-0.025	-0.058	-0.145	-0.101^{1}	0.074	0.029	0.083
P4Coll75	-0.123^{1}	-0.033	-0.097	-0.167	-0.171^{3}	0.054	0.021	0.077
P4Coll90	-0.124^{1}	-0.039	-0.128	-0.181	-0.184^{3}	0.039	0.014	0.071

Tableau C.2 – Variables d'UPs et difficulté : ρ de Spearman pour les deux corpus et les 4 seuils. Le taux de significativité des corrélations est indiqué comme suit : $^{1} : < 0.05$; $^{2} : < 0.01$; $^{3} : < 0.001$.

ANNEXE D

RÉSULTATS DES PROCÉDURES D'OPTIMISATION DES MÉTA-PARAMÈTRES

Ces annexes comprennent les résultats des différentes procédures d'optimisation pour la « baseline », les modèles *Expert1*, *Expert2* et par sélection automatique de variables.

D.1 La « baseline »

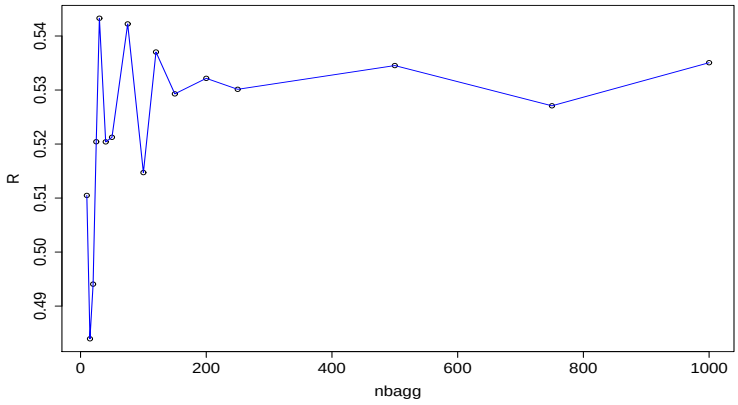
D.1.1 Les arbres de décision

Performances des modèles « baseline » par arbre de décision en fonction des paramètres cp et $minsplit$, estimées par validation croisée à dix échantillons sur *Dev6CE* :

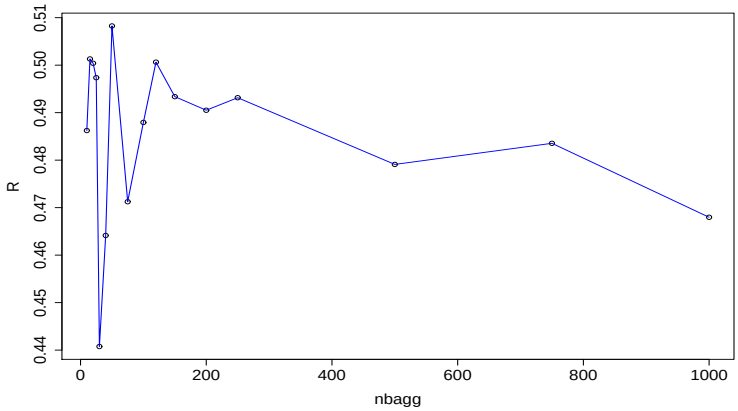
cp	$minsplit$	R	$acc-cont$	cp	$minsplit$	R	$acc-cont$
$0,1$	5	$0,592$	$55,2$	$0,1$	15	$0,591$	$55,2$
$0,05$	5	$0,595$	60	$0,05$	15	$0,611$	57
$0,01$	5	$0,618$	67	$0,01$	15	$0,643$	$68,3$
$0,005$	5	$0,619$	$67,8$	$0,005$	15	$0,623$	$66,1$
$0,001$	5	$0,62$	70	$0,001$	15	$0,612$	$68,3$
$0,0005$	5	$0,595$	$66,1$	$0,0005$	15	$0,623$	$67,4$
$0,1$	10	$0,592$	55	$0,1$	20	$0,592$	$55,2$
$0,05$	10	$0,60$	$56,5$	$0,05$	20	$0,585$	$56,5$
$0,01$	10	$0,618$	$68,69$	$0,01$	20	$0,627$	$69,1$
$0,005$	10	$0,624$	70	$0,005$	20	$0,626$	$64,8$
$0,001$	10	$0,605$	$72,6$	$0,001$	20	$0,627$	$67,8$
$0,0005$	10	$0,611$	66	$0,0005$	20	$0,627$	$69,6$

D.1.2 Le bagging

Performances des modèles « baseline » par bagging en fonction du nombre d'itération, estimées par validation croisée à dix échantillons sur *Dev6CE* :

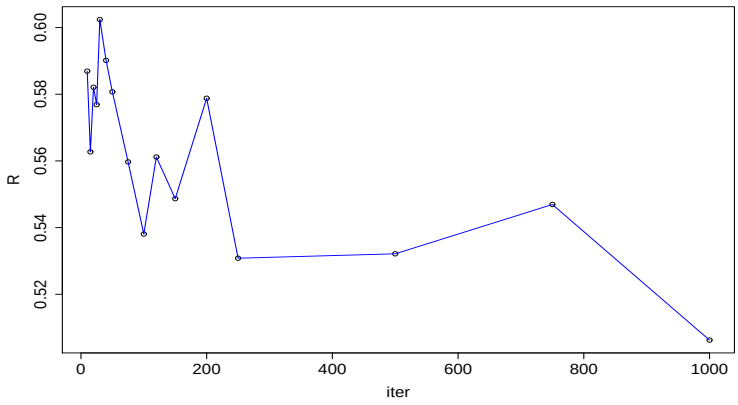


Performances des modèles « baseline » par bagging en fonction du nombre d'itération, estimées par validation croisée à dix échantillons sur *Dev9CE* :

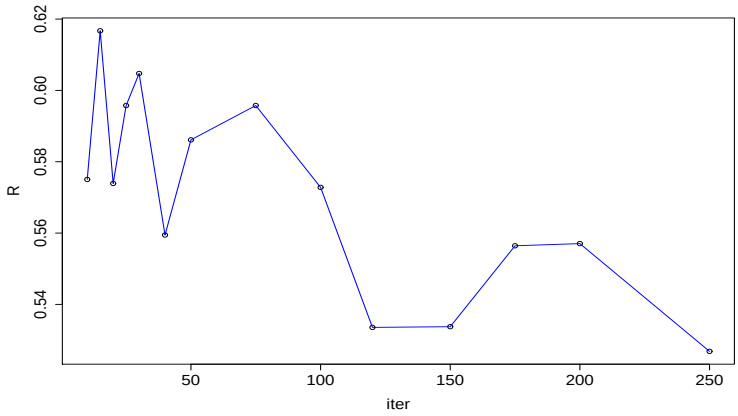


D.1.3 Le boosting

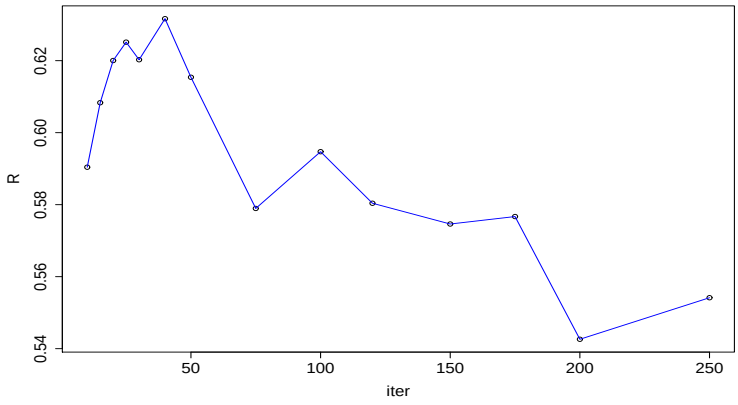
Performances des modèles « baseline » par boosting (*stump*) en fonction du nombre d'itération, estimées par validation croisée à dix échantillons sur *Dev6CE* :



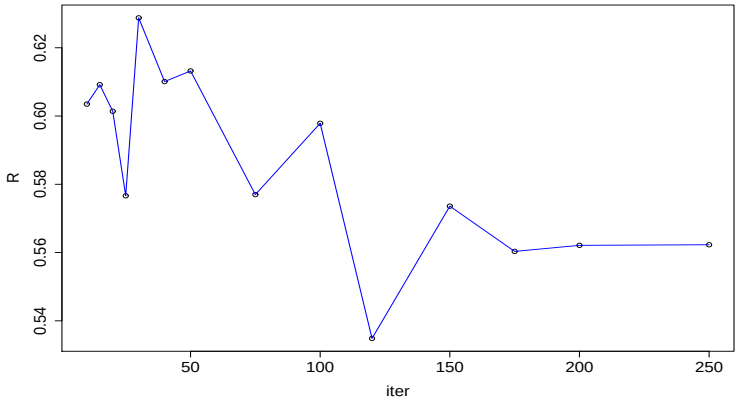
Performances des modèles « baseline » par boosting (*four*) en fonction du nombre d'itération, estimées par validation croisée à dix échantillons sur *Dev6CE* :



Performances des modèles « baseline » par boosting (*stump*) en fonction du nombre d'itération, estimées par validation croisée à dix échantillons sur *Dev9CE* :



Performances des modèles « baseline » par boosting (*four*) en fonction du nombre d'itération, estimées par validation croisée à dix échantillons sur *Dev9CE* :



D.1.4 Les SVM

Performances des modèles « baseline » par SVM en fonction des paramètres $gamma$ et C, estimées par validation croisée à dix échantillons sur *Dev6CE* :

$gamma$	C	$acc-cont$	$gamma$	C	$acc-cont$
$1e-04$	1	54.78261	$1e-04$	25	62.6087
$5e-04$	1	54.34783	$5e-04$	25	65.65217
0.001	1	54.78261	0.001	25	67.82609
0.002	1	60.43478	0.002	25	68.69565
0.004	1	66.95652	0.004	25	66.52174
0.006	1	67.3913	0.006	25	66.95652
0.01	1	67.3913	0.01	25	68.69565
0.1	1	66.95652	0.1	25	66.95652
0.5	1	66.95652	0.5	25	70.86957
$1e-04$	2	54.78261	$1e-04$	40	66.52174
$5e-04$	2	54.78261	$5e-04$	40	67.82609
0.001	2	60.43478	0.001	40	68.26087
0.002	2	66.08696	0.002	40	67.82609
0.004	2	66.52174	0.004	40	66.52174
0.006	2	65.65217	0.006	40	69.13043
0.01	2	65.65217	0.01	40	68.69565
0.1	2	66.52174	0.1	40	67.82609
0.5	2	68.69565	0.5	40	69.13043
$1e-04$	5	54.78261	$1e-04$	50	67.3913
$5e-04$	5	62.17391	$5e-04$	50	67.82609
0.001	5	66.95652	0.001	50	68.69565
0.002	5	68.26087	0.002	50	66.95652
0.004	5	66.52174	0.004	50	67.82609
0.006	5	67.3913	0.006	50	69.56522
0.01	5	68.26087	0.01	50	68.69565
0.1	5	70	0.1	50	67.3913

0.5	5	70	0.5	50	69.56522
$1e-04$	8	54.78261	$1e-04$	75	65.65217
$5e-04$	8	66.08696	$5e-04$	75	67.82609
0.001	8	66.08696	0.001	75	69.13043
0.002	8	66.52174	0.002	75	66.95652
0.004	8	65.65217	0.004	75	69.13043
0.006	8	68.69565	0.006	75	70.86957
0.01	8	68.26087	0.01	75	66.52174
0.1	8	68.69565	0.1	75	67.82609
0.5	8	68.69565	0.5	75	68.69565
$1e-04$	10	55.21739	$1e-04$	100	68.26087
$5e-04$	10	66.95652	$5e-04$	100	68.69565
0.001	10	68.26087	0.001	100	66.95652
0.002	10	67.3913	0.002	100	67.3913
0.004	10	67.82609	0.004	100	70.43478
0.006	10	67.3913	0.006	100	68.69565
0.01	10	66.52174	0.01	100	67.3913
0.1	10	68.69565	0.1	100	67.3913
0.5	10	69.13043	0.5	100	68.26087
$1e-04$	12	55.21739	$1e-04$	500	68.69565
$5e-04$	12	65.65217	$5e-04$	500	69.13043
0.001	12	65.65217	0.001	500	69.13043
0.002	12	67.82609	0.002	500	69.56522
0.004	12	68.69565	0.004	500	66.08696
0.006	12	67.82609	0.006	500	67.3913
0.01	12	66.52174	0.01	500	69.56522
0.1	12	68.26087	0.1	500	66.95652
0.5	12	68.69565	0.5	500	64.78261
$1e-04$	15	59.13043	$1e-04$	1000	66.95652
$5e-04$	15	66.08696	$5e-04$	1000	69.13043
0.001	15	65.65217	0.001	1000	69.13043
0.002	15	67.3913	0.002	1000	67.82609
0.004	15	67.3913	0.004	1000	67.3913
0.006	15	66.08696	0.006	1000	69.13043
0.01	15	67.3913	0.01	1000	68.69565
0.1	15	67.82609	0.1	1000	66.95652
0.5	15	69.56522	0.5	1000	65.65217
$1e-04$	20	60.86957			
$5e-04$	20	68.26087			
0.001	20	67.82609			
0.002	20	68.26087			
0.004	20	67.82609			
0.006	20	67.3913			
0.01	20	68.26087			
0.1	20	68.26087			
0.5	20	70			

Performances des modèles « baseline » par SVM en fonction des paramètres $gamma$ et C, estimées par validation croisée à dix échantillons sur *Dev9CE* :

$gamma$	C	$acc-cont$	$gamma$	C	$acc-cont$
$1e-04$	1	40.80882	$1e-04$	25	40.80882
$5e-04$	1	40.80882	$5e-04$	25	50.75
0.001	1	40.80882	0.001	25	49.07353
0.002	1	40.80882	0.002	25	53.64706
0.004	1	43.45588	0.004	25	54.23529
0.006	1	50.08824	0.006	25	55.11765
0.01	1	49.11765	0.01	25	54.86765
0.1	1	53.30882	0.1	25	54.91176
0.5	1	53.94118	0.5	25	52.64706
$1e-04$	2	40.80882	$1e-04$	40	44.83824
$5e-04$	2	40.80882	$5e-04$	40	50.20588
0.001	2	40.80882	0.001	40	53.05882
0.002	2	43.45588	0.002	40	54.23529
0.004	2	48.36765	0.004	40	55.70588
0.006	2	49.91176	0.006	40	55.16176
0.01	2	51	0.01	40	56.04412
0.1	2	55.66176	0.1	40	55.95588
0.5	2	54.23529	0.5	40	52.64706
$1e-04$	5	40.80882	$1e-04$	50	51.26471
$5e-04$	5	40.80882	$5e-04$	50	49.07353
0.001	5	50.97059	0.001	50	53.64706
0.002	5	48.82353	0.002	50	53.94118
0.004	5	50.75	0.004	50	54.82353
0.006	5	49.91176	0.006	50	55.16176
0.01	5	53.35294	0.01	50	55.45588
0.1	5	54.07353	0.1	50	54.52941
0.5	5	54.27941	0.5	50	51.47059
$1e-04$	8	40.80882	$1e-04$	75	47.82353
$5e-04$	8	44.83824	$5e-04$	75	51.33824
0.001	8	48.36765	0.001	75	54.82353
0.002	8	50.45588	0.002	75	55.11765
0.004	8	50.79412	0.004	75	54.86765
0.006	8	53.94118	0.006	75	55.45588
0.01	8	54.82353	0.01	75	55.41176
0.1	8	54.36765	0.1	75	54.73529
0.5	8	53.73529	0.5	75	51.47059
$1e-04$	10	40.80882	$1e-04$	100	49.11765
$5e-04$	10	51.26471	$5e-04$	100	53.64706
0.001	10	48.82353	0.001	100	53.94118
0.002	10	50.75	0.002	100	55.16176
0.004	10	52.47059	0.004	100	55.45588
0.006	10	53.35294	0.006	100	55.75
0.01	10	55.11765	0.01	100	56.83824
0.1	10	55.16176	0.1	100	54.52941
0.5	10	53.77941	0.5	100	52.10294
$1e-04$	12	40.80882	$1e-04$	500	53.64706
$5e-04$	12	50.04412	$5e-04$	500	54.61765
0.001	12	50.5	0.001	500	55.16176
0.002	12	48.44118	0.002	500	56.04412
0.004	12	53.94118	0.004	500	55.70588
0.006	12	54.82353	0.006	500	56.72059
0.01	12	54.23529	0.01	500	54.36765
0.1	12	55.11765	0.1	500	52.80882

0.5	12	52.89706	0.5	500	51.30882
$1e-04$	15	40.80882	$1e-04$	1000	53.94118
$5e-04$	15	47.82353	$5e-04$	1000	54.91176
0.001	15	50.75	0.001	1000	54.86765
0.002	15	49.02941	0.002	1000	55.45588
0.004	15	53.39706	0.004	1000	57.80882
0.006	15	54.52941	0.006	1000	55.29412
0.01	15	55.70588	0.01	1000	54.27941
0.1	15	55.45588	0.1	1000	52.76471
0.5	15	52.60294	0.5	1000	51.01471
$1e-04$	20	40.80882			
$5e-04$	20	48.82353			
0.001	20	50.20588			
0.002	20	53.05882			
0.004	20	54.23529			
0.006	20	54.52941			
0.01	20	55.16176			
0.1	20	55.75			
0.5	20	52.35294			

D.2 Les modèles *Expert1*

D.2.1 Les arbres de décision

Performances des modèles *Expert1* par arbre de décision en fonction des paramètres cp et $minsplit$, estimées par validation croisée à dix échantillons sur *Dev6CE* :

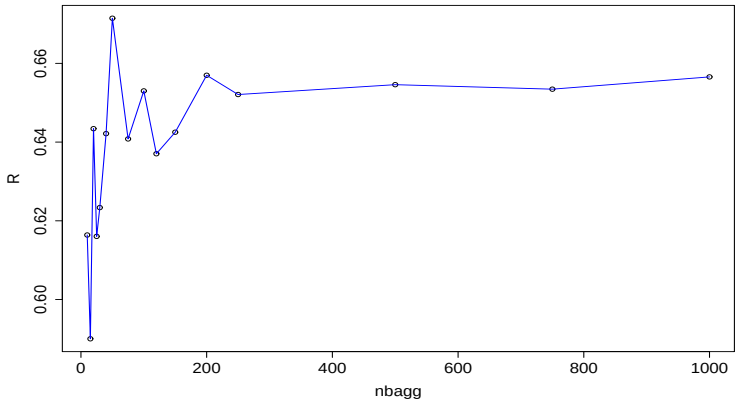
cp	$minsplit$	$acc-cont$	cp	$minsplit$	$acc-cont$
$0,1$	2	$55,2$	$0,1$	10	$55,2$
$0,05$	2	67	$0,05$	10	$64,8$
$0,01$	2	$68,7$	$0,01$	10	$71,7$
$0,005$	2	$71,3$	$0,005$	10	$70,4$
$0,001$	2	$70,4$	$0,001$	10	$69,6$
$0,0005$	2	$69,6$	$0,0005$	10	$69,6$
$0,1$	5	$55,2$	$0,1$	15	$55,2$
$0,05$	5	$67,8$	$0,05$	15	$66,1$
$0,01$	5	$70,4$	$0,01$	15	$71,3$
$0,005$	5	$68,3$	$0,005$	15	$68,7$
$0,001$	5	$70,9$	$0,001$	15	$70,4$
$0,0005$	5	$70,4$	$0,0005$	15	$70,4$
$0,1$	20	$55,2$	$0,005$	20	$69,6$
$0,05$	20	$64,3$	$0,001$	20	$70,4$
$0,01$	20	$69,6$	$0,0005$	20	$70,4$

Performances des modèles *Expert1* par arbre de décision en fonction des paramètres cp et $minsplit$, estimées par validation croisée à dix échantillons sur *Dev9CE* :

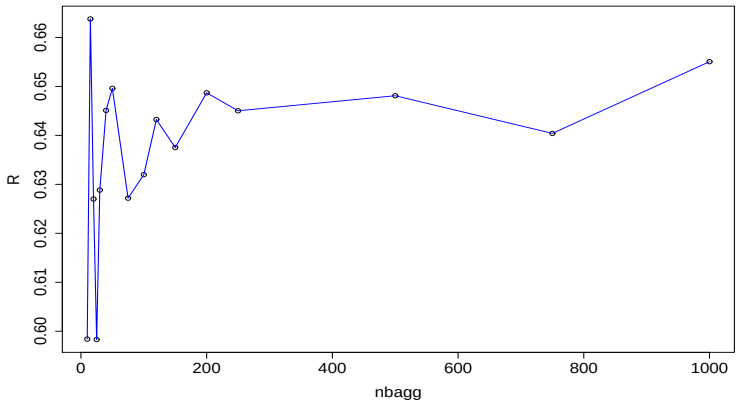
cp	$minsplit$	$acc-cont$	cp	$minsplit$	$acc-cont$
$0,1$	2	$41,6$	$0,1$	10	$44,8$
$0,05$	2	$48,1$	$0,05$	10	$47,5$
$0,01$	2	$49,1$	$0,01$	10	$48,6$
$0,005$	2	$50,5$	$0,005$	10	$49,7$
$0,001$	2	$50,9$	$0,001$	10	$51,2$
$0,0005$	2	$52,1$	$0,0005$	10	$49,7$
$0,1$	5	$41,6$	$0,1$	15	$44,8$
$0,05$	5	$47,8$	$0,05$	15	$47,5$
$0,01$	5	$48,3$	$0,01$	15	$49,1$
$0,005$	5	$48,3$	$0,005$	15	$48,8$
$0,001$	5	$52,7$	$0,001$	15	$48,9$
$0,0005$	5	52	$0,0005$	15	$49,7$
$0,1$	20	$44,8$	$0,005$	20	$48,1$
$0,05$	20	$48,4$	$0,001$	20	52
$0,01$	20	50	$0,0005$	20	50

D.2.2 Le bagging

Performances des modèles *Expert1* par bagging en fonction du nombre d'itération, estimées par validation croisée à dix échantillons sur *Dev6CE* :

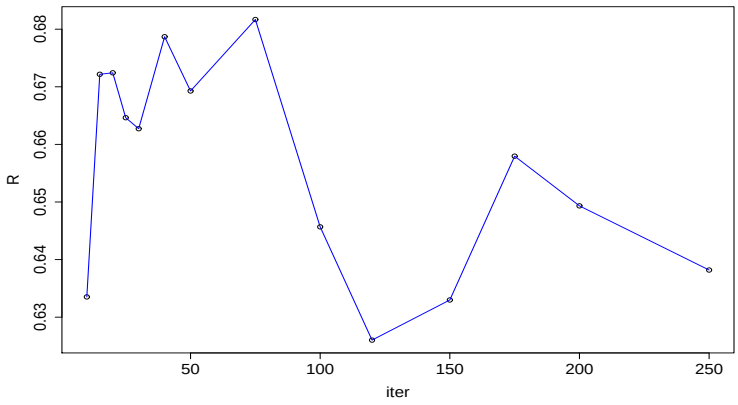


Performances des modèles *Expert1* par bagging en fonction du nombre d'itération, estimées par validation croisée à dix échantillons sur *Dev9CE* :

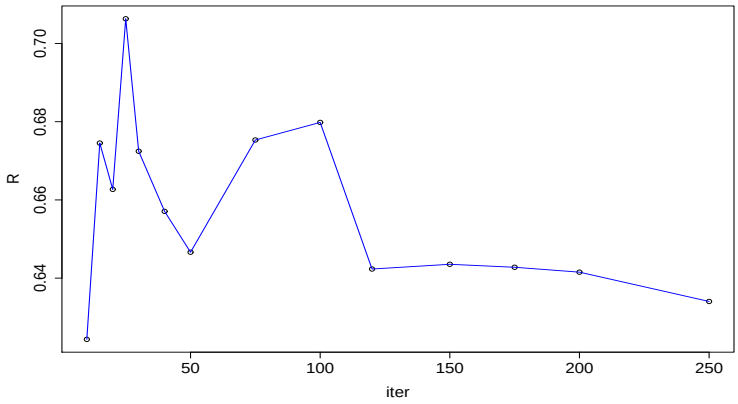


D.2.3 Le boosting

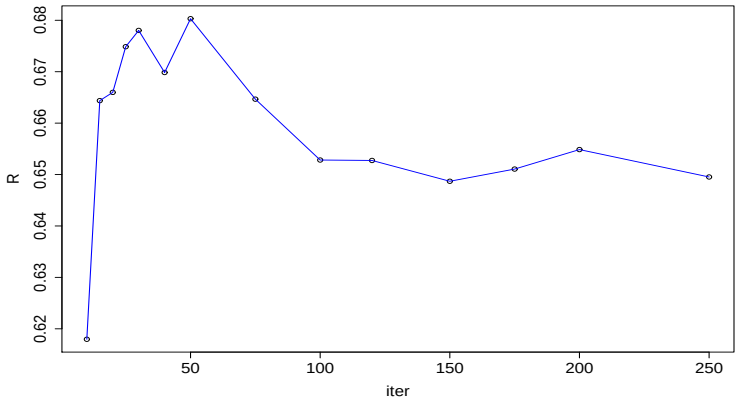
Performances des modèles *Expert1* par boosting (*stump*) en fonction du nombre d'itération, estimées par validation croisée à dix échantillons sur *Dev6CE* :



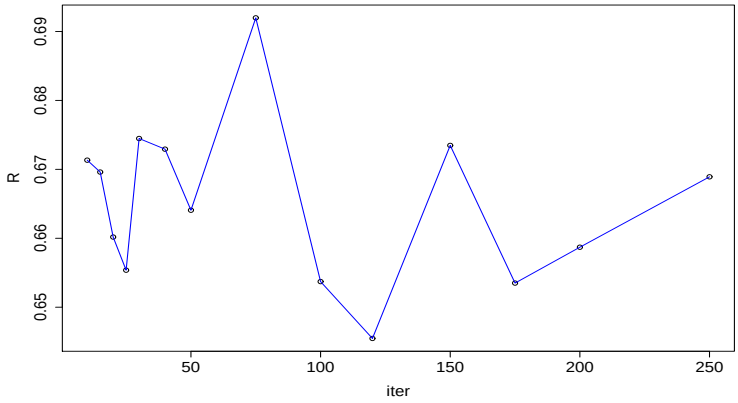
Performances des modèles *Expert1* par boosting (*four*) en fonction du nombre d'itération, estimées par validation croisée à dix échantillons sur *Dev6CE* :



Performances des modèles *Expert1* par boosting (*stump*) en fonction du nombre d'itération, estimées par validation croisée à dix échantillons sur *Dev9CE* :



Performances des modèles *Expert1* par boosting (*four*) en fonction du nombre d'itération, estimées par validation croisée à dix échantillons sur *Dev9CE* :



D.2.4 Les SVM

Performances des modèles *Expert1* par SVM en fonction des paramètres $gamma$ et C, estimées par validation croisée à dix échantillons sur *Dev6CE* :

$gamma$	C	$acc-cont$	$gamma$	C	$acc-cont$
$1e-04$	1	54.66403	$1e-04$	25	67.27273
$5e-04$	1	54.66403	$5e-04$	25	69.01186
0.001	1	56.40316	0.001	25	68.16206
0.002	1	65.5336	0.002	25	69.01186
0.004	1	69.01186	0.004	25	68.12253
0.006	1	71.18577	0.006	25	68.55731
0.01	1	69.44664	0.01	25	68.9921
0.1	1	66.77866	0.1	25	69.4664
0.5	1	69.88142	0.5	25	68.083
$1e-04$	2	54.66403	$1e-04$	40	68.57708
$5e-04$	2	56.83794	$5e-04$	40	69.48617
0.001	2	65.5336	0.001	40	69.01186
0.002	2	68.57708	0.002	40	68.12253
0.004	2	69.88142	0.004	40	68.55731
0.006	2	69.01186	0.006	40	68.12253
0.01	2	69.92095	0.01	40	68.9921
0.1	2	67.64822	0.1	40	69.88142
0.5	2	69.03162	0.5	40	66.32411
$1e-04$	5	54.66403	$1e-04$	50	70.75099
$5e-04$	5	67.27273	$5e-04$	50	68.16206
0.001	5	70.75099	0.001	50	69.01186
0.002	5	70.3162	0.002	50	68.57708
0.004	5	69.48617	0.004	50	68.9921
0.006	5	68.59684	0.006	50	68.12253
0.01	5	69.42688	0.01	50	69.42688
0.1	5	68.55731	0.1	50	70.3162

0.5	5	70.35573	0.5	50	63.71542
$1e-04$	8	55.09881	$1e-04$	75	69.88142
$5e-04$	8	68.57708	$5e-04$	75	68.57708
0.001	8	69.01186	0.001	75	68.55731
0.002	8	68.57708	0.002	75	68.57708
0.004	8	68.57708	0.004	75	68.12253
0.006	8	68.55731	0.006	75	69.42688
0.01	8	67.68775	0.01	75	68.9921
0.1	8	68.95257	0.1	75	69.90119
0.5	8	69.01186	0.5	75	61.08696
$1e-04$	10	56.83794	$1e-04$	100	70.3162
$5e-04$	10	70.3162	$5e-04$	100	69.01186
0.001	10	70.3162	0.001	100	68.57708
0.002	10	69.92095	0.002	100	69.01186
0.004	10	69.01186	0.004	100	68.55731
0.006	10	68.12253	0.006	100	70.29644
0.01	10	68.55731	0.01	100	68.12253
0.1	10	67.2332	0.1	100	69.44664
0.5	10	69.42688	0.5	100	62.3913
$1e-04$	12	61.62055	$1e-04$	500	68.57708
$5e-04$	12	70.3162	$5e-04$	500	69.01186
0.001	12	69.01186	0.001	500	69.01186
0.002	12	67.70751	0.002	500	69.42688
0.004	12	69.44664	0.004	500	67.68775
0.006	12	68.55731	0.006	500	68.10277
0.01	12	68.9921	0.01	500	70.73123
0.1	12	67.66798	0.1	500	69.48617
0.5	12	67.2332	0.5	500	61.50198
$1e-04$	15	67.27273	$1e-04$	1000	68.57708
$5e-04$	15	69.88142	$5e-04$	1000	69.44664
0.001	15	69.90119	0.001	1000	68.9921
0.002	15	68.59684	0.002	1000	68.55731
0.004	15	68.12253	0.004	1000	67.2332
0.006	15	67.68775	0.006	1000	71.18577
0.01	15	68.55731	0.01	1000	69.44664
0.1	15	67.25296	0.1	1000	69.48617
0.5	15	67.19368	0.5	1000	62.82609
$1e-04$	20	65.96838			
$5e-04$	20	70.3162			
0.001	20	69.92095			
0.002	20	69.01186			
0.004	20	68.55731			
0.006	20	69.44664			
0.01	20	68.12253			
0.1	20	68.57708			
0.5	20	68.04348			

Performances des modèles *Expert1* par SVM en fonction des paramètres $gamma$ et C, estimées par validation croisée à dix échantillons sur *Dev9CE* :

$gamma$	C	$acc-cont$	$gamma$	C	$acc-cont$
$1e-04$	1	44	$1e-04$	25	52.10294
$5e-04$	1	44	$5e-04$	25	52.13235
0.001	1	44	0.001	25	52.42647
0.002	1	48.36765	0.002	25	53.60294
0.004	1	51.22059	0.004	25	55.07353
0.006	1	51.80882	0.006	25	54.77941
0.01	1	53.19118	0.01	25	54.48529
0.1	1	55.66176	0.1	25	59.05882
0.5	1	58.13235	0.5	25	51.58824
$1e-04$	2	44	$1e-04$	40	51.51471
$5e-04$	2	44	$5e-04$	40	53.30882
0.001	2	48.66176	0.001	40	54.19118
0.002	2	51.51471	0.002	40	54.19118
0.004	2	53.19118	0.004	40	54.82353
0.006	2	52.51471	0.006	40	53.89706
0.01	2	53.60294	0.01	40	57.17647
0.1	2	55.66176	0.1	40	59.05882
0.5	2	60.77941	0.5	40	49.86765
$1e-04$	5	44	$1e-04$	50	50.63235
$5e-04$	5	52.10294	$5e-04$	50	52.42647
0.001	5	50.63235	0.001	50	53.89706
0.002	5	53.94118	0.002	50	54.48529
0.004	5	53.01471	0.004	50	54.48529
0.006	5	52.47059	0.006	50	55.07353
0.01	5	54.77941	0.01	50	56.83824
0.1	5	58.22059	0.1	50	58.47059
0.5	5	53.76471	0.5	50	50.45588
$1e-04$	8	44	$1e-04$	75	52.60294
$5e-04$	8	51.51471	$5e-04$	75	52.72059
0.001	8	52.60294	0.001	75	54.19118
0.002	8	52.42647	0.002	75	54.77941
0.004	8	53.30882	0.004	75	54.82353
0.006	8	54.19118	0.006	75	56.29412
0.01	8	54.48529	0.01	75	56.58824
0.1	8	60.57353	0.1	75	56.66176
0.5	8	52	0.5	75	50.41176
$1e-04$	10	44	$1e-04$	100	54.23529
$5e-04$	10	50.63235	$5e-04$	100	53.89706
0.001	10	54.23529	0.001	100	54.48529
0.002	10	53.30882	0.002	100	54.19118
0.004	10	53.89706	0.004	100	55.75
0.006	10	54.19118	0.006	100	56.29412
0.01	10	54.73529	0.01	100	58.30882
0.1	10	60.32353	0.1	100	56.11765
0.5	10	53.22059	0.5	100	50.41176
$1e-04$	12	44	$1e-04$	500	53.89706
$5e-04$	12	50.63235	$5e-04$	500	53.64706
0.001	12	52.72059	0.001	500	55.11765
0.002	12	52.72059	0.002	500	56.58824
0.004	12	53.89706	0.004	500	56.88235
0.006	12	55.07353	0.006	500	57.72059
0.01	12	55.32353	0.01	500	58.80882
0.1	12	61.16176	0.1	500	50.07353

0.5	12	53.26471	0.5	500	50.92647
$1e-04$	15	44	$1e-04$	1000	54.23529
$5e-04$	15	52.89706	$5e-04$	1000	55.11765
0.001	15	51.54412	0.001	1000	55.11765
0.002	15	52.76471	0.002	1000	56
0.004	15	54.48529	0.004	1000	57.67647
0.006	15	55.07353	0.006	1000	59.10294
0.01	15	54.44118	0.01	1000	59.05882
0.1	15	60.27941	0.1	1000	50.75
0.5	15	53.26471	0.5	1000	52.39706
$1e-04$	20	48.36765			
$5e-04$	20	54.23529			
0.001	20	53.30882			
0.002	20	53.89706			
0.004	20	54.77941			
0.006	20	55.02941			
0.01	20	53.60294			
0.1	20	59.64706			
0.5	20	52.42647			

D.3 Les modèles *Expert2*

D.3.1 Les arbres de décision

Performances des modèles *Expert2* par arbre de décision en fonction des paramètres cp et $minsplit$, estimées par validation croisée à dix échantillons sur *Dev6CE* :

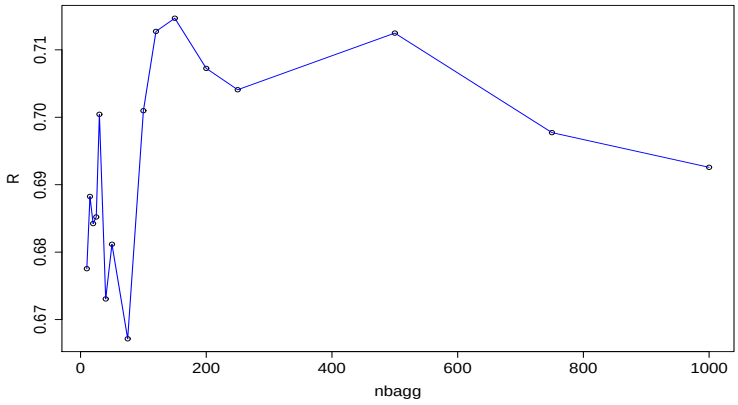
cp	$minsplit$	$acc-cont$	cp	$minsplit$	$acc-cont$
0,1	2	55,7	0,1	10	55,7
0,05	2	61,3	0,05	10	63,9
0,01	2	67,4	0,01	10	70,4
0,005	2	67,8	0,005	10	65,7
0,001	2	66,5	0,001	10	68,3
0,0005	2	67,8	0,0005	10	67,4
0,1	5	55,6	0,1	15	55,7
0,05	5	61,3	0,05	15	63,5
0,01	5	66,5	0,01	15	67,4
0,005	5	65,2	0,005	15	68,7
0,001	5	66,5	0,001	15	66,5
0,0005	5	63	0,0005	15	67,8
0,1	20	55,7	0,005	20	66,5
0,05	20	65,7	0,001	20	66,5
0,01	20	64,8	0,0005	20	66,5

Performances des modèles *Expert2* par arbre de décision en fonction des paramètres cp et $minsplit$, estimées par validation croisée à dix échantillons sur *Dev9CE* :

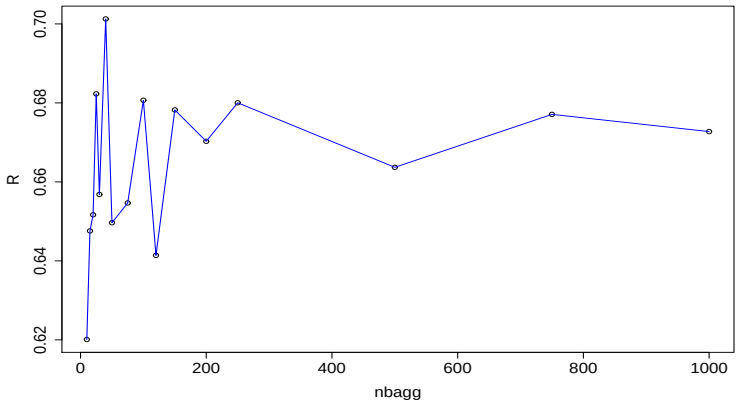
cp	$minsplit$	$acc-cont$	cp	$minsplit$	$acc-cont$
0,1	2	44,8	0,1	10	44,8
0,05	2	47,2	0,05	10	47,2
0,01	2	49,9	0,01	10	49,9
0,005	2	50,8	0,005	10	51,1
0,001	2	50,8	0,001	10	49,7
0,0005	2	51,1	0,0005	10	50,2
0,1	5	44,8	0,1	15	44,8
0,05	5	47,2	0,05	15	46
0,01	5	49,9	0,01	15	51,1
0,005	5	51,1	0,005	15	50,8
0,001	5	51,1	0,001	15	50,8
0,0005	5	50,8	0,0005	15	50,8
0,1	20	44,8	0,005	20	53,2
0,05	20	47,2	0,001	20	50,8
0,01	20	51,7	0,0005	20	52,4

D.3.2 Le bagging

Performances des modèles *Expert2* par bagging en fonction du nombre d'itération, estimées par validation croisée à dix échantillons sur *Dev6CE* :

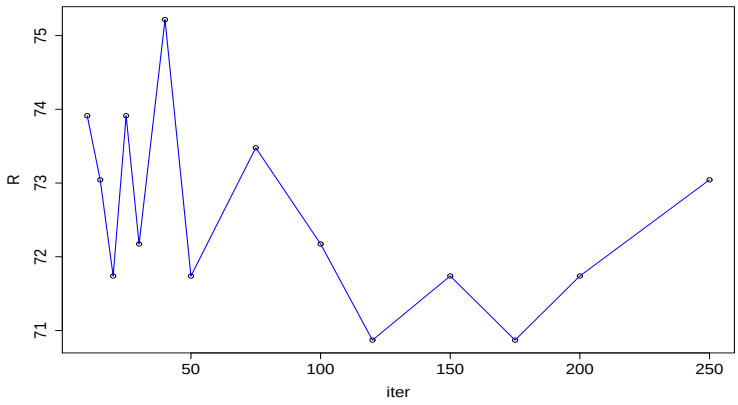


Performances des modèles *Expert2* par bagging en fonction du nombre d'itération, estimées par validation croisée à dix échantillons sur *Dev9CE* :

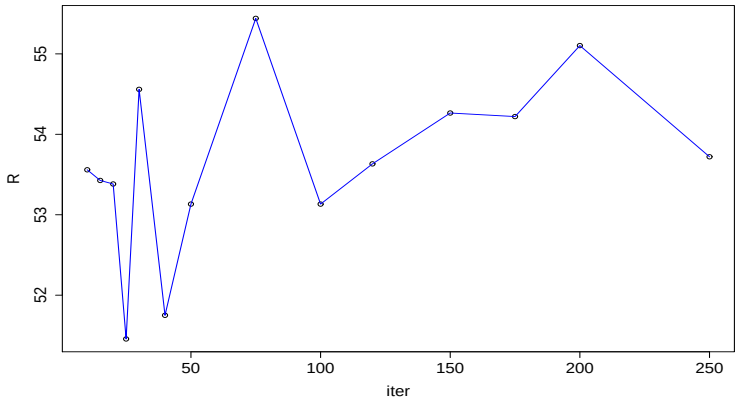


D.3.3 Le boosting

Performances des modèles *Expert2* par boosting (*four*) en fonction du nombre d'itération, estimées par validation croisée à dix échantillons sur *Dev6CE* :



Performances des modèles *Expert1* par boosting (*four*) en fonction du nombre d'itération, estimées par validation croisée à dix échantillons sur *Dev9CE* :



D.3.4 Les SVM

Performances des modèles *Expert2* par SVM en fonction des paramètres $gamma$ et C, estimées par validation croisée à dix échantillons sur *Dev6CE* :

$gamma$	C	$acc-cont$	$gamma$	C	$acc-cont$
$1e-04$	1	49.36759	$1e-04$	25	65.05929
$5e-04$	1	49.36759	$5e-04$	25	66.81818
0.001	1	58.51779	0.001	25	69.88142
0.002	1	63.32016	0.002	25	71.18577
0.004	1	64.16996	0.004	25	70.75099
0.006	1	66.79842	0.006	25	71.62055
0.01	1	66.79842	0.01	25	69.88142
0.1	1	70.3162	0.1	25	65.92885
0.5	1	67.70751	0.5	25	70.77075

$1e-04$	2	49.36759	$1e-04$	40	63.73518
$5e-04$	2	59.38735	$5e-04$	40	68.12253
0.001	2	63.75494	0.001	40	70.75099
0.002	2	64.16996	0.002	40	69.44664
0.004	2	66.79842	0.004	40	70.75099
0.006	2	67.25296	0.006	40	69.88142
0.01	2	67.70751	0.01	40	67.27273
0.1	2	68.57708	0.1	40	66.36364
0.5	2	66.77866	0.5	40	71.64032
$1e-04$	5	49.80237	$1e-04$	50	65.49407
$5e-04$	5	64.6245	$5e-04$	50	70.3162
0.001	5	65.49407	0.001	50	70.75099
0.002	5	66.79842	0.002	50	71.62055
0.004	5	69.01186	0.004	50	69.88142
0.006	5	69.44664	0.006	50	69.88142
0.01	5	71.62055	0.01	50	67.27273
0.1	5	68.9921	0.1	50	64.18972
0.5	5	65.92885	0.5	50	70.75099
$1e-04$	8	53.28063	$1e-04$	75	67.25296
$5e-04$	8	63.73518	$5e-04$	75	70.75099
0.001	8	68.10277	0.001	75	69.44664
0.002	8	67.66798	0.002	75	72.0751
0.004	8	69.88142	0.004	75	71.18577
0.006	8	72.05534	0.006	75	67.70751
0.01	8	69.01186	0.01	75	68.16206
0.1	8	68.9921	0.1	75	65.05929
0.5	8	65.94862	0.5	75	70.75099
$1e-04$	10	59.38735	$1e-04$	100	65.92885
$5e-04$	10	65.49407	$5e-04$	100	70.75099
0.001	10	66.36364	0.001	100	71.62055
0.002	10	68.12253	0.002	100	69.88142
0.004	10	71.62055	0.004	100	69.88142
0.006	10	70.3162	0.006	100	69.03162
0.01	10	70.3162	0.01	100	68.9921
0.1	10	68.9921	0.1	100	65.05929
0.5	10	66.3834	0.5	100	70.75099
$1e-04$	12	59.8419	$1e-04$	500	71.18577
$5e-04$	12	66.3834	$5e-04$	500	71.14625
0.001	12	67.25296	0.001	500	69.88142
0.002	12	69.4664	0.002	500	70.33597
0.004	12	70.75099	0.004	500	68.55731
0.006	12	69.44664	0.006	500	66.3834
0.01	12	70.3162	0.01	500	65.92885
0.1	12	68.9921	0.1	500	61.62055
0.5	12	67.27273	0.5	500	70.75099
$1e-04$	15	62.01581	$1e-04$	1000	71.62055
$5e-04$	15	67.25296	$5e-04$	1000	69.86166
0.001	15	67.68775	0.001	1000	70.75099
0.002	15	69.88142	0.002	1000	70.33597
0.004	15	69.88142	0.004	1000	67.25296
0.006	15	69.44664	0.006	1000	65.92885
0.01	15	70.3162	0.01	1000	64.60474
0.1	15	67.25296	0.1	1000	62.45059
0.5	15	68.6166	0.5	1000	70.75099

$1e-04$	20	64.18972			
$5e-04$	20	66.36364			
0.001	20	68.12253			
0.002	20	71.62055			
0.004	20	69.88142			
0.006	20	71.64032			
0.01	20	70.75099			
0.1	20	65.92885			
0.5	20	69.4664			

Performances des modèles *Expert2* par SVM en fonction des paramètres $gamma$ et C, estimées par validation croisée à dix échantillons sur *Dev9CE* :

$gamma$	C	$acc-cont$	$gamma$	C	$acc-cont$
$1e-04$	1	41.73529	$1e-04$	25	47.27941
$5e-04$	1	41.73529	$5e-04$	25	56.83824
0.001	1	42.02941	0.001	25	57.63235
0.002	1	48.41176	0.002	25	57.58824
0.004	1	49.45588	0.004	25	59.7353
0.006	1	52.10294	0.006	25	59.77941
0.01	1	56.29412	0.01	25	58.05882
0.1	1	59.2647	0.1	25	54.32353
0.5	1	55.86765	0.5	25	51.51471
$1e-04$	2	41.73529	$1e-04$	40	50.04412
$5e-04$	2	42.02941	$5e-04$	40	59.10294
0.001	2	48.41176	0.001	40	57.58824
0.002	2	49.16176	0.002	40	60.2353
0.004	2	54.52941	0.004	40	60.32353
0.006	2	58.05882	0.006	40	58.35294
0.01	2	59.44118	0.01	40	58.69118
0.1	2	58.17647	0.1	40	56
0.5	2	55.32353	0.5	40	52.98529
$1e-04$	5	42.02941	$1e-04$	50	50.33824
$5e-04$	5	47.57353	$5e-04$	50	57.63235
0.001	5	50.08824	0.001	50	57.29412
0.002	5	56.25	0.002	50	60.02941
0.004	5	58.55882	0.004	50	58.60294
0.006	5	58.80882	0.006	50	58.64706
0.01	5	57.54412	0.01	50	58.9853
0.1	5	57.58824	0.1	50	55.66176
0.5	5	53.94118	0.5	50	52.98529
$1e-04$	8	42.02941	$1e-04$	75	54.23529
$5e-04$	8	49.75	$5e-04$	75	57.58824
0.001	8	56.29412	0.001	75	59.64706
0.002	8	57.92647	0.002	75	60.07353
0.004	8	58.51471	0.004	75	58.64706
0.006	8	57.58824	0.006	75	58.14706
0.01	8	59.64706	0.01	75	59.77941
0.1	8	58.47059	0.1	75	56.29412

0.5	8	52.35294	0.5	75	53.57353
$1e-04$	10	42.02941	$1e-04$	100	56.83824
$5e-04$	10	50.33824	$5e-04$	100	57.58824
0.001	10	56.54412	0.001	100	59.7353
0.002	10	58.55882	0.002	100	59.19118
0.004	10	57.08824	0.004	100	57.22059
0.006	10	58.76471	0.006	100	58.94118
0.01	10	59.10294	0.01	100	58.85294
0.1	10	56.45588	0.1	100	55.75
0.5	10	53.52941	0.5	100	53.86765
$1e-04$	12	44.08824	$1e-04$	500	57.83824
$5e-04$	12	52.05882	$5e-04$	500	59.86765
0.001	12	57.72059	0.001	500	58.94118
0.002	12	58.22059	0.002	500	58.39706
0.004	12	57.29412	0.004	500	58.30882
0.006	12	58.80882	0.006	500	58.89706
0.01	12	59.7353	0.01	500	56.54412
0.1	12	56.5	0.1	500	54.02941
0.5	12	52.64706	0.5	500	53.86765
$1e-04$	15	45.85294	$1e-04$	1000	60.32353
$5e-04$	15	54.52941	$5e-04$	1000	57.76471
0.001	15	57.67647	0.001	1000	59.19118
0.002	15	58.22059	0.002	1000	59.52941
0.004	15	58.76471	0.004	1000	58.85294
0.006	15	59.94118	0.006	1000	58.30882
0.01	15	58.89706	0.01	1000	53.60294
0.1	15	55.95588	0.1	1000	52.76471
0.5	15	52.64706	0.5	1000	53.86765
$1e-04$	20	48.41176			
$5e-04$	20	56.54412			
0.001	20	59.10294			
0.002	20	57.33824			
0.004	20	59.64706			
0.006	20	60.61765			
0.01	20	58.64706			
0.1	20	56			
0.5	20	52.10294			

D.4 Les modèles par sélection automatique de variables

D.4.1 Les arbres de décision

Performances des modèles par sélection automatique de variables reposant sur les arbres de décision, en fonction des paramètres cp et $minsplit$, estimées par validation croisée à dix échantillons sur *Dev6CE* :

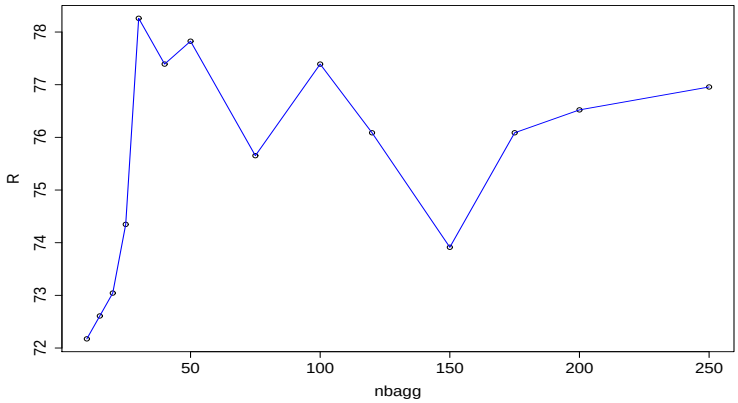
cp	$minsplit$	$acc-cont$	cp	$minsplit$	$acc-cont$
$0,1$	2	$56,1$	$0,1$	10	$56,1$
$0,05$	2	$63,5$	$0,05$	10	$63,9$
$0,01$	2	$69,6$	$0,01$	10	$65,2$
$90,005$	2	$68,7$	$0,005$	10	$68,7$
$0,001$	2	$71,7$	$0,001$	10	$71,3$
$0,0005$	2	$71,3$	$0,0005$	10	$65,7$
$0,1$	5	$56,1$	$0,1$	15	$56,1$
$0,05$	5	$63,5$	$0,05$	15	$63,5$
$0,01$	5	$65,7$	$0,01$	15	$70,9$
$0,005$	5	$67,8$	$0,005$	15	$70,9$
$0,001$	5	$69,1$	$0,001$	15	$66,5$
$0,0005$	5	$71,3$	$0,0005$	15	$69,6$
$0,1$	20	$56,1$	$0,005$	20	$70,9$
$0,05$	20	$60,9$	$0,001$	20	$70,9$
$0,01$	20	$70,4$	$0,0005$	20	$65,7$

Performances des modèles par sélection automatique de variables reposant sur les arbres de décision, en fonction des paramètres cp et $minsplit$, estimées par validation croisée à dix échantillons sur *Dev9CE* :

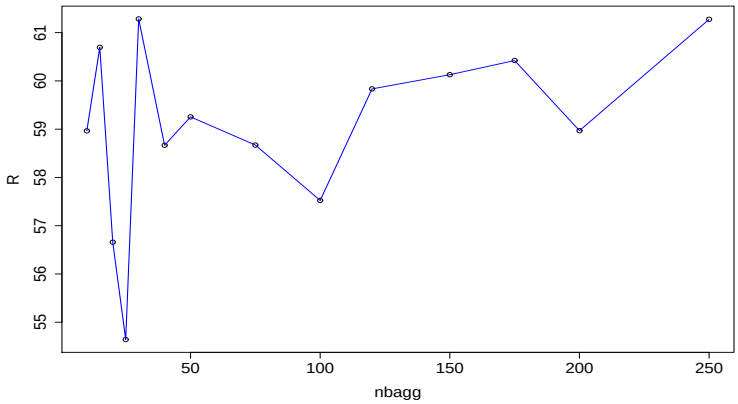
cp	$minsplit$	$acc-cont$	cp	$minsplit$	$acc-cont$
$0,1$	2	$46,5$	$0,1$	10	$46,5$
$0,05$	2	$49,2$	$0,05$	10	$45,7$
$0,01$	2	$53,8$	$0,01$	10	$54,1$
$0,005$	2	$53,8$	$0,005$	10	$53,8$
$0,001$	2	$53,5$	$0,001$	10	$52,4$
$0,0005$	2	$53,8$	$0,0005$	10	$52,6$
$0,1$	5	$46,5$	$0,1$	15	$46,5$
$0,05$	5	$47,7$	$0,05$	15	$47,7$
$0,01$	5	55	$0,01$	15	$54,4$
$0,005$	5	$55,3$	$0,005$	15	$53,2$
$0,001$	5	$52,9$	$0,001$	15	$53,2$
$0,0005$	5	$52,9$	$0,0005$	15	55
$0,1$	20	$46,5$	$0,005$	20	55
$0,05$	20	$47,7$	$0,001$	20	$57,6$
$0,01$	20	$52,1$	$0,0005$	20	$53,8$

D.4.2 Le bagging

Performances des modèles par sélection automatique des variables reposant sur le bagging, en fonction du nombre d'itération, estimées par validation croisée à dix échantillons sur *Dev6CE* :



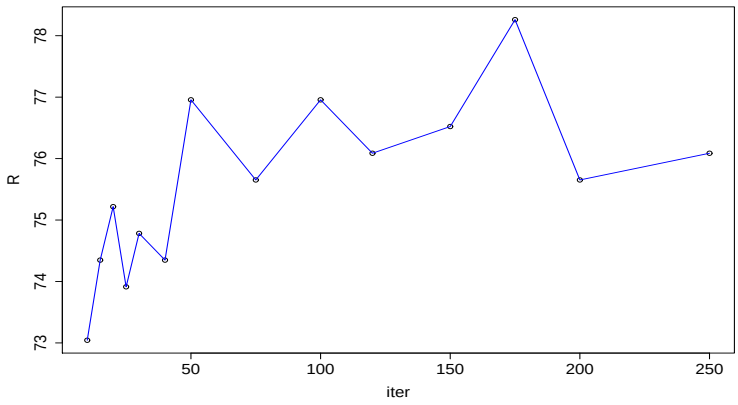
Performances des modèles *Expert2* par sélection automatique des variables reposant sur le bagging, en fonction du nombre d'itération, estimées par validation croisée à dix échantillons sur *Dev9CE* :



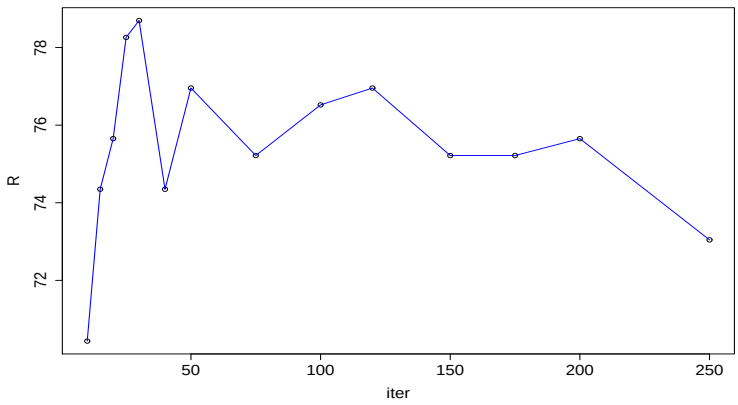
D.4.3 Le boosting

Performances des modèles par sélection automatique des variables reposant sur le boosting (*stump*) en fonction du nombre d'itération, estimées par validation croisée à dix échantillons

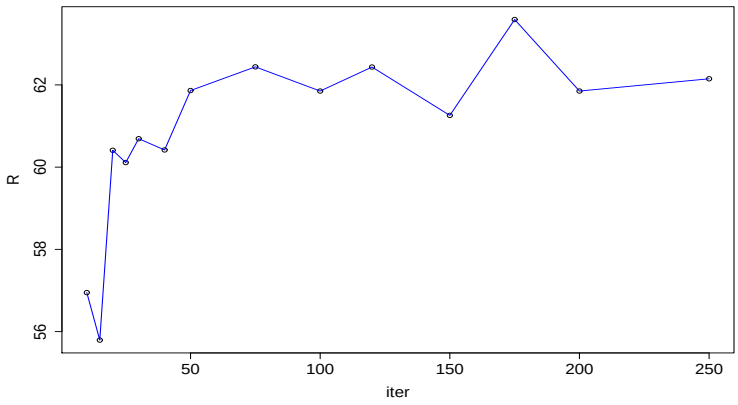
sur *Dev6CE* :



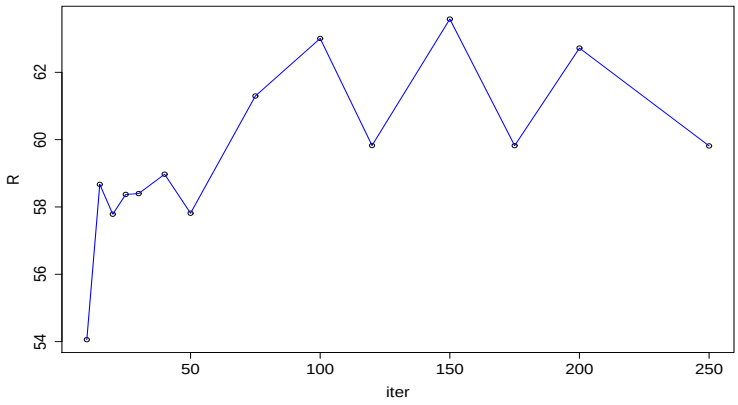
Performances des modèles par sélection automatique des variables reposant sur le boosting (*four*) en fonction du nombre d'itération, estimées par validation croisée à dix échantillons sur *Dev6CE* :



Performances des modèles par sélection automatique des variables reposant sur le boosting (*stump*) en fonction du nombre d'itération, estimées par validation croisée à dix échantillons sur *Dev9CE* :



Performances des modèles par sélection automatique des variables reposant sur le boosting (*four*) en fonction du nombre d'itération, estimées par validation croisée à dix échantillons sur *Dev9CE* :



D.4.4 Les SVM

Performances des modèles par sélection automatique des variables reposant sur des SVM, en fonction des paramètres $gamma$ et C, estimées par validation croisée à dix échantillons sur *Dev6CE* :

$gamma$	C	$acc-cont$	$gamma$	C	$acc-cont$
$1e-04$	1	55.02415	$1e-04$	25	70.72464
$5e-04$	1	67.66184	$5e-04$	25	75.14976
0.001	1	69.41063	0.001	25	74.24155
0.002	1	71.64251	0.002	25	74.24155
0.004	1	72.92754	0.004	25	73.81643
0.006	1	73.78744	0.006	25	72.95652
0.01	1	74.22222	0.01	25	73.40097

0.1	1	72.92754	0.1	25	75.11111
0.5	1	51.93237	0.5	25	48.47343
$1e-04$	2	57.19807	$1e-04$	40	73.37198
$5e-04$	2	69.84541	$5e-04$	40	76.00966
0.001	2	71.19807	0.001	40	74.24155
0.002	2	72.9372	0.002	40	73.37198
0.004	2	74.22222	0.004	40	71.64251
0.006	2	75.5459	0.006	40	72.52174
0.01	2	74.24155	0.01	40	73.3913
0.1	2	75.11111	0.1	40	75.11111
0.5	2	48.47343	0.5	40	48.47343
$1e-04$	5	67.66184	$1e-04$	50	72.49275
$5e-04$	5	70.29952	$5e-04$	50	74.24155
0.001	5	72.92754	0.001	50	75.55556
0.002	5	75.13043	0.002	50	72.06763
0.004	5	76.87923	0.004	50	72.51208
0.006	5	75.5459	0.006	50	73.83575
0.01	5	74.24155	0.01	50	72.95652
0.1	5	75.11111	0.1	50	75.11111
0.5	5	48.47343	0.5	50	48.47343
$1e-04$	8	68.96618	$1e-04$	75	74.25121
$5e-04$	8	73.37198	$5e-04$	75	74.68599
0.001	8	74.25121	0.001	75	74.24155
0.002	8	75.57488	0.002	75	72.9372
0.004	8	75.5459	0.004	75	72.95652
0.006	8	74.24155	0.006	75	73.3913
0.01	8	74.26087	0.01	75	73.83575
0.1	8	75.11111	0.1	75	75.11111
0.5	8	48.47343	0.5	75	48.47343
$1e-04$	10	69.84541	$1e-04$	100	75.1401
$5e-04$	10	72.49275	$5e-04$	100	75.12077
0.001	10	75.58454	0.001	100	73.37198
0.002	10	76.44444	0.002	100	69.4686
0.004	10	74.23188	0.004	100	72.06763
0.006	10	75.98068	0.006	100	74.26087
0.01	10	72.95652	0.01	100	73.83575
0.1	10	75.11111	0.1	100	75.11111
0.5	10	48.47343	0.5	100	48.47343
$1e-04$	12	72.48309	$1e-04$	500	75.55556
$5e-04$	12	73.37198	$5e-04$	500	72.52174
0.001	12	74.70531	0.001	500	72.9372
0.002	12	75.56522	0.002	500	72.52174
0.004	12	75.5459	0.004	500	73.3913
0.006	12	75.12077	0.006	500	74.26087
0.01	12	72.95652	0.01	500	73.83575
0.1	12	75.11111	0.1	500	75.11111
0.5	12	48.47343	0.5	500	48.47343
$1e-04$	15	71.1884	$1e-04$	1000	72.49275
$5e-04$	15	74.68599	$5e-04$	1000	73.81643
0.001	15	75.58454	0.001	1000	72.92754
0.002	15	74.67633	0.002	1000	73.40097
0.004	15	75.5459	0.004	1000	73.3913
0.006	15	73.38164	0.006	1000	74.26087
0.01	15	74.27053	0.01	1000	73.83575

0.1	15	75.11111	0.1	1000	75.11111
0.5	15	48.47343	0.5	1000	48.47343
$1e-04$	20	69.8744			
$5e-04$	20	75.58454			
0.001	20	75.1401			
0.002	20	73.80676			
0.004	20	75.99034			
0.006	20	71.65217			
0.01	20	74.2802			
0.1	20	75.11111			
0.5	20	48.47343			

Performances des modèles par sélection automatique des variables reposant sur des SVM, en fonction des paramètres $gamma$ et C, estimées par validation croisée à dix échantillons sur *Dev9CE* :

$gamma$	C	$acc-cont$	$gamma$	C	$acc-cont$
$1e-04$	1	39.5942	$1e-04$	25	57.80124
$5e-04$	1	44.80331	$5e-04$	25	59.84265
0.001	1	50.30642	0.001	25	62.43478
0.002	1	55.78468	0.002	25	64.4472
0.004	1	55.7764	0.004	25	62.7205
0.006	1	58.9648	0.006	25	61.27122
0.01	1	58.9648	0.01	25	60.99793
0.1	1	58.10766	0.1	25	58.69979
0.5	1	36.41822	0.5	25	36.13251
$1e-04$	2	40.17391	$1e-04$	40	56.93582
$5e-04$	2	51.4617	$5e-04$	40	60.4058
0.001	2	57.80124	0.001	40	63.8675
0.002	2	57.22567	0.002	40	63.57764
0.004	2	60.12836	0.004	40	61.26708
0.006	2	60.70393	0.006	40	60.42236
0.01	2	60.98965	0.01	40	59.84265
0.1	2	58.69979	0.1	40	58.69979
0.5	2	36.13251	0.5	40	36.13251
$1e-04$	5	45.08903	$1e-04$	50	56.9441
$5e-04$	5	57.7971	$5e-04$	50	61.85921
0.001	5	56.9441	0.001	50	63.57764
0.002	5	60.13665	0.002	50	64.15735
0.004	5	62.72464	0.004	50	61.27122
0.006	5	64.46791	0.006	50	59.56108
0.01	5	62.14907	0.01	50	59.5528
0.1	5	58.69979	0.1	50	58.69979
0.5	5	36.13251	0.5	50	36.13251
$1e-04$	8	50.29814	$1e-04$	75	59.55694
$5e-04$	8	56.93582	$5e-04$	75	63.29193
0.001	8	59.55694	0.001	75	63.28778
0.002	8	60.12422	0.002	75	62.14079
0.004	8	64.46377	0.004	75	60.71222

0.006	8	62.14493	0.006	75	61.28364
0.01	8	63.88406	0.01	75	59.5528
0.1	8	58.69979	0.1	75	58.69979
0.5	8	36.13251	0.5	75	36.13251
$1e-04$	10	50.59627	$1e-04$	100	60.1325
$5e-04$	10	57.22981	$5e-04$	100	63.28778
0.001	10	60.1325	0.001	100	62.71222
0.002	10	61.27536	0.002	100	61.27122
0.004	10	62.43064	0.004	100	61.00621
0.006	10	63.87578	0.006	100	60.70393
0.01	10	63.01863	0.01	100	60.12422
0.1	10	58.69979	0.1	100	58.69979
0.5	10	36.13251	0.5	100	36.13251
$1e-04$	12	52.04141	$1e-04$	500	62.70393
$5e-04$	12	57.52795	$5e-04$	500	63.2795
0.001	12	60.12422	0.001	500	61.85093
0.002	12	63.02277	0.002	500	61.28778
0.004	12	62.13665	0.004	500	59.83023
0.006	12	63.87164	0.006	500	60.4058
0.01	12	62.43892	0.01	500	60.12422
0.1	12	58.69979	0.1	500	58.69979
0.5	12	36.13251	0.5	500	36.13251
$1e-04$	15	52.32298	$1e-04$	1000	61.85507
$5e-04$	15	59.26708	$5e-04$	1000	62.4265
0.001	15	61.28778	0.001	1000	62.72464
0.002	15	63.30435	0.002	1000	62.43478
0.004	15	64.7412	0.004	1000	59.83023
0.006	15	62.7205	0.006	1000	60.4058
0.01	15	61.28364	0.01	1000	60.12422
0.1	15	58.69979	0.1	1000	58.69979
0.5	15	36.13251	0.5	1000	36.13251
$1e-04$	20	57.51139			
$5e-04$	20	59.5528			
0.001	20	60.1118			
0.002	20	63.58178			
0.004	20	63.30021			
0.006	20	61.56108			
0.01	20	61.57764			
0.1	20	58.69979			
0.5	20	36.13251			

Bibliographie

ABRY, D., DAÂS, Y., FERT, C., DESCHAMPS, H., RICHAUD, F. et SPERANDIO, C. (2008a). *Ici 2*. CLE International, Paris.

ABRY, D., FERT, C., PARPETTE, C., STAUBER, J., SORIA, M. et BORG, S. (2008b). *Ici 1*. CLE International, Paris.

AGRESTI, A. (2002). *Categorical Data Analysis. 2nd edition*. Wiley-Interscience, New York.

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. *In Second international symposium on information theory*, volume 1, pages 267–281. Springer Verlag.

ALDERSON, J. (1979). The cloze procedure and proficiency in english as a foreign language. *TESOL quarterly*, 13(2):219–227.

ALDERSON, J. (1984). Reading in a foreign language : a reading problem or a language problem ? *In* ALDERSON, J. et URQUHART, A., éditeurs : *Reading in a Foreign Language*, pages 1–24. Longman, New York.

ALDERSON, J. (2000). *Assessing reading*. Cambridge University Press, Cambridge.

ALDERSON, J. (2007). The cefr and the need for more research. *The Modern Language Journal*, 91(4):659–663.

ALDERSON, J. et URQUHART, A. (1984). Introduction : What is reading. *In* ALDERSON, J. et URQUHART, A., éditeurs : *Reading in a Foreign Language*, pages xv–xxviii. Longman, New York.

ALLEN, J. et WIDDOWSON, H. (1974). Teaching the Communicative Use of English. *International Review of Applied Linguistics in Language Teaching*, 12(1):1–21.

ALLINGTON, R. et FLEMING, J. (1978). The misreading of high-frequency words. *The Journal of Special Education*, 12(4):417–421.

ALUISIO, S., SPECIA, L., GASPERIN, C. et SCARTON, C. (2010). Readability assessment for text simplification. *In Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9, Los Angeles.

ALZOLA, C. et HARRELL, F. (2002). An introduction to s and the hmisc and design libraries.

AMIRAN, M. et JONES, B. (1982). Toward a new definition of readability. *Educational Psychologist*, 17(1):13–30.

ANAGNOSTOU, N. et WEIR, G. (2006). From corpus-based collocation frequencies to readability measure. *ICT in the Analysis, Teaching and Learning of Languages*, pages 33–46.

ANDERSON, J. (1981). Analysing the Readability of English and non-English Texts in the Classroom with Lix. *In Paper presented at the Australian Reading Association Conference*, Darwin.

ANDERSON, J. (1983). Lix and Rix : variations on a little-known readability index. *Journal of Reading*, 26(6):490–496.

ANDERSON, R. et PEARSON, P. (1984). A schema-theoretic view of basic processes in reading comprehension. *In* ANDERSON, C. et PEARSON, P., éditeurs : *Handbook of reading research*, pages 255–291. Longman, New York.

ANDERSON, R. et PICHERT, J. (1978). Recall of previously unrecallable information following a shift in perspective. *Journal of Verbal Learning and Verbal Behavior*, 17(1):1–12.

ANDREWS, S. (1997). The effect of orthographic similarity on lexical retrieval : Resolving neighborhood conflicts. *Psychonomic Bulletin & Review*, 4(4):439–461.

ANTONIADIS, G. et GRUSSON, Y. (1996). Modélisation et génération automatique de la lisibilité de textes. *In ILN 96 : Informatique et Langue Naturelle*.

ANTONIADIS, G. et PONTON, C. (2004). MIRTO : un système au service de l'enseignement des langues. *In Proc. of UNTELE 2004*, Compiègne, France.

AQUINO, M., MOSBERG, L. et SHARRON, M. (1969). Reading comprehension difficulty as a function of content area and linguistic complexity. *The Journal of Experimental Educational*, 37(4):1–4.

ARMBRUSTER, B. (1984). The problem of "Inconsiderate text". *In* DUFFEY, G., éditeur : *Compehension instruction : Perspectives and suggestions*, pages 202–217. Longman, New York.

ARNOLD, H. (1932). A List of Graded Vocabularies and a Method of Grading. *Modern Language Journal*, 16(8):644–655.

ATKINSON, R. et SHIFFRIN, R. (1968). Human memory : A proposed system and its control processes. *In* SPENCE, K. et SPENCE, J., éditeurs : *The Psychology of Learning and Motivation*, volume 2, pages 89–195. Academic Press, New York.

AUGÉ, H., Cañada PUJOLS, M., MARLHENS, C. et MARTIN, L. (2005a). *Tout va bien ! 1*. CLE International, Paris.

AUGÉ, H., Cañada PUJOLS, M., MARLHENS, C. et MARTIN, L. (2005b). *Tout va bien ! 2*. CLE International, Paris.

AUGÉ, H., Cañada PUJOLS, M., MARLHENS, C. et MARTIN, L. (2006). *Tout va bien ! 3*. CLE International, Paris.

BAAYEN, R., PIEPENBROCK, R. et van RIJN, H. (1993). *The {CELEX} lexical data base on {CD-ROM}*. Linguistic Data Consortium, Philadelphia : University of Pennsylvania.

BADDELEY, A. (2003). Working memory : looking back and looking forward. *Nature Reviews : Neuroscience*, 4(10):829–839.

BAHNS, J. et ELDAW, M. (1993). Should We Teach EFL Students Collocations ? *System*, 21(1):101–14.

BALDWIN, R. et KAUFMAN, R. (1979). A concurrent validity study of the Raygor readability estimate. *Journal of Reading*, 23(2):148–153.

BALDWIN, T., BANNARD, C., TANAKA, T. et WIDDOWS, D. (2003). An empirical model of multiword expression decomposability. *In Proceedings of the ACL 2003 workshop on Multiword expressions : analysis, acquisition and treatment*, pages 89–96.

BALOTA, D. et CHUMBLEY, J. (1984). Are lexical decisions a good measure of lexical access ? : The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology*, 10(3):340–357.

BARBARESI, A. (2011). La complexité linguistique, méthode d'analyse.

BARBER, H. et KUTAS, M. (2007). Interplay between computational models and cognitive electrophysiology in visual word recognition. *Brain Research Reviews*, 53(1):98–123.

BARKER, D. et STOKES, W. (1968). A Simplification of the Revised Lorge Readability Formula. *The Journal of Educational Research*, 61(9):398–400.

BARZILAY, R. et LAPATA, M. (2005). Modeling local coherence : An entity-based approach. pages 141–148.

BARZILAY, R. et LAPATA, M. (2008). Modeling local coherence : An entity-based approach. *Computational Linguistics*, 34(1):1–34.

BEAUDET, C. (2001). Clarté, lisibilité, intelligibilité des textes : un état de la question et une proposition pédagogique. *Recherches en rédaction professionnelle*, 1(1).

BECKER, C. (1976). Allocation of attention during visual word recognition. *Journal of Experimental Psychology : Human Perception and Performance*, 2(4):556–566.

BELL, T. (2001). Extensive reading : Speed and comprehension. *The Reading Matrix*, 1(1).

BENOÎT, J.-P. (1986a). Lisibilité dans la littérature enfantine : quels textes ? *Les Actes de lecture*, 15.

BENOÎT, J.-P. (1986b). Revue critique des formules de lisibilité, 60 ans de formules de lisibilité : qu'en reste-t-il ? *Pratiques*, 52:45–63.

BENOÎT, J.-P. (1986c). Vous avez dit lisibilité ? *Les Actes de lecture*, 16.

BEREITER, C. (1991). Implications of connectionism for thinking about rules. *Educational Researcher*, 20(3):10–16.

BERNHARDT, E. (1991). A psycholinguistic perspective on second language literacy. *Reading in two languages*, 8:31–44.

BERNHARDT, E. (2003). Challenges to reading research from a multilingual world. *Reading Research Quarterly*, 37(1):112–117.

BERNHARDT, E. (2005). Progress and procrastination in second language reading. *Annual Review of Applied Linguistics*, 25:133–150.

BERNHARDT, E. et KAMIL, M. (2006). Second-language reading. *In* BROWN, K., éditeur : *Encyclopedia of Languages and Linguistics. 2nd Edition*, pages 88–95. Elsevier, Oxford.

BERTHET, A., HUGOT, C., KIZIRIAN, V., SAMPSONIS, B. et WAENDENDRIES, M. (2006a). *Alter Ego 1*. Hachette, Paris.

BERTHET, A., HUGOT, C., KIZIRIAN, V., SAMPSONIS, B. et WAENDENDRIES, M. (2006b). *Alter Ego 2*. Hachette, Paris.

BESSE, H. (1980). Enseigner la compétence de communication ? *Le Français dans le Monde*, 153:41–47.

BESSE, H. et GALISSON, R. (1980). *Polémique en didactique : du renouveau en question*. CLE international, Paris.

BESSE, P. (2005). Data mining II - Modélisation Statistique et Apprentissage. Rapport technique, Laboratoire de statistiques et de probabilités, Université Paul Sabatier.

BESTGEN, Y. (2011). Lsvma : au plus deux composants pour apparier des résumés à des articles. *In Actes DEFT 2011*, pages 105–114.

BINDER, J., WESTBURY, C., MCKIERNAN, K., POSSING, E. et MEDLER, D. (2005). Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience*, 17(6):905–917.

BINKLEY, M. (1988). New ways of assessing text difficulty. *In* ZAKALUK, B. et SAMUELS, S., éditeurs : *Readability : Its Past, Present and Future*, pages 98–120. International Reading Association, Newark, Delaware.

BISHOP, C. (2006). *Pattern recognition and machine learning*. Springer, New York.

BJÖRNSSON, C. (1983). Readability of newspapers in 11 languages. *Reading Research Quarterly*, 18(4):480–497.

BLAIR, I., URLAND, G. et MA, J. (2002). Using Internet search engines to estimate word frequency. *Behavior Research Methods, Instruments, & Computers*, 34(2):286–290.

BLAU, E. (1982). The effect of syntax on readability for ESL students in Puerto Rico. *TESOL Quarterly*, 16(4):517–528.

BLEASDALE, F. (1987). Concreteness-dependent associative priming : Separate lexical organization for concrete and abstract words. *Journal of Experimental Psychology : Learning, Memory, and Cognition*, 13(4):582–594.

BLOOMER, R. (1956). Word Length and Complexity Variables in Spelling Difficulty. *The Journal of Educational Research*, 49(7):531–535.

BLOOMER, R. (1959). Level of Abstraction as a Function of Modifer Load. *The Journal of Educational Research*, 52(7):269–272.

BLOOMFIELD, A. et DAILL, E. (2006). *DELF B2, 200 activités*. CLE International, Paris.

BODER, D. (1940). The adjective-verb quotient : A contribution to the psychology of language. *Psychological Record*, 3:310–343.

BONIN, P., CHALARD, M., MÉOT, A. et FAYOL, M. (2001). Age-of acquisition and word frequency in the lexical decision task : Further evidence from the French language. *Current Psychology of Cognition*, 20(6):401–444.

BORMUTH, J. (1966). Readability : A new approach. *Reading research quarterly*, 1(3):79–132.

BORMUTH, J. (1969a). *Development of Readability Analysis*. Rapport technique, Projet n˚7-0052, U.S. Office of Education, Bureau of Research, Department of Health, Education and Welfare, Washington, DC.

BORMUTH, J. (1969b). Factor validity of cloze tests as measures of reading comprehension ability. *Reading Research Quarterly*, 4(3):358–365.

BORMUTH, J., MANNING, J., CARR, J. et PEARSON, D. (1970). Children's comprehension of between-and within-sentence syntactic structures. *Journal of educational psychology*, 61(5):349–357.

BOROWSKY, R. et MASSON, M. (1996). Semantic ambiguity effects in word identification. *Journal of Experimental Psychology : Learning, Memory, and Cognition*, 22(1):63.

BOSER, B., GUYON, I. et VAPNIK, V. (1992). A training algorithm for optimal margin classifiers. *In Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152.

BOTEL, M. et GRANOWSKY, A. (1972). A Formula for Measuring Syntactic Complexity ; A Directional Effort. *Elementary English*, 49(4):513–16.

BOUMA, H. (1971). Visual recognition of isolated lower-case letters. *Vision Research*, 11(5):459–474.

BOWERS, J. (2000). In defense of abstractionist theories of repetition priming and word identification. *Psychonomic bulletin and review*, 7(1):83–99.

BOYER, J. (1992). La lisibilité. *Revue française de pédagogie*, 99(1):5–14.

BRAMER, M. (2007). *Principles of Data Mining*. Springer-Verlag, London.

BREIMAN, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

BREIMAN, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

BREIMAN, L., FRIEDMAN, H., OLSEN, R. et STONE, J. (1984). *Classification and regression trees*. Chapman & Hall, New York.

BREWER, W. (1972). Is reading a letter-by-letter process ? *In* KAVANAGH, J. F. et MATTINGLY, I. G., éditeurs : *Language by ear and by eye*, pages 359–365. MIT Press, Cambridge.

BRITTON, B. et GÜLGÖZ, S. (1991). Using kintsch's computational model to improve instructional text : Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology*, 83(3):329–345.

BROWN, C., SNODGRASS, T., KEMPER, S., HERMAN, R. et COVINGTON, M. (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior research methods*, 40(2):540–545.

BROWN, J. (1952). The Flesch Formula 'Through the Looking Glass'. *College English*, 13(7):393–394.

BROWN, M. et FORSYTHE, A. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, 69(346):364–367.

BROWN, W. (1965). Science textbook selection and the Dale-Chall formula. *School Science and Mathematics*, 65(2):164–167.

BRUCE, B., RUBIN, A. et STARR, K. (1981). *Why readability formulas fail*. Rapport technique, Reading Education Report No. 28, University of Illinois at Urbana-Champaign, Champaign, IL.

BRYSBAERT, M. (1996). Word frequency affects naming latency in Dutch when age of acquisition is controlled. *European Journal of Cognitive Psychology*, 8(2):185–194.

BRYSBAERT, M., LANGE, M. et VAN WIJNENDAELE, I. (2000). The effects of age-of-acquisition and frequency-of-occurrence in visual word recognition : Further evidence from the Dutch language. *European Journal of Cognitive Psychology*, 12(1):65–85.

BRYSBAERT, M. et VITU, F. (1998). Word skipping : Implications for theories of eye movement control in reading. *In* UNDERWOOD, G., éditeur : *Eye guidance in reading and scene perception*, pages 125–148. Elsevier, Oxford.

CABELLO, B. (1984). Cultural interference in reading comprehension : An alternative explanation. *Bilingual Review*, 11(1):12–20.

CAMPBELL, F. et WURTZ, R. (1978). Saccadic omission : why we do not see a gray-out during a saccadic eye movement. *Vision Research*, 18:1297–1303.

CAPELLE, G. et MENAND, R. (2003). *Taxi ! 1*. Hachette, Paris.

CAPUCHO, F., DENYER, M., LABASCOULE, J. et ROYER, C. (2007). *Rond Point 3*. Difusión, Barcelone.

CARAMAZZA, A. et BRONES, I. (1979). Lexical access in bilinguals. *Bulletin of the Psychonomic Society*, 13:212–214.

CARREIRAS, M., CARRIEDO, N., ALONSO, M. et FERNÁNDEZ, A. (1997). The role of verb tense and verb aspect in the foregrounding of information during reading. *Memory & Cognition*, 25(4):438–446.

CARRELL, P. (1982). Cohesion is not coherence. *TESOL quarterly*, 16(4):479–488.

CARRELL, P. (1983). Three components of background knowledge in reading comprehension. *Language Learning*, 33(2):183–203.

CARRELL, P. (1984). The effects of rhetorical organization on esl readers. *TESOL quarterly*, 18(3):441–469.

CARRELL, P. (1987). Readability in ESL. *Reading in a Foreign Language*, 4(1):21–40.

CARRELL, P. (1989). SLA and Classroom Instruction : Reading. *Annual Review of Applied Linguistics*, 9:223–242.

CARRELL, P., DEVINE, J. et ESKEY, D., éditeurs (1989). *Interactive approaches to second language reading*. Cambridge Univ. Press, Cambridge.

CARTERETTE, E. et JONES, M. (1963). Redundancy in children's texts. *Science*, 140(3573): 1309–1311.

CARVER, R. (1974a). *Improving Reading Comprehension : Measuring Readability*. Rapport technique, Final Report, Contract No. N00014-72-C0240. American Institues for Research in the Behavioral Sciences, Silver Spring, Maryland.

CARVER, R. (1974b). Measuring the primary effect of reading : Reading-storage technique, understanding judgements, and clozeunderstanding, information processing, and learning from prose materials. *Journal of Reading Behavior*, 6:249–274.

CARVER, R. (1975). Measuring prose difficulty using the Rauding Scale. *Reading Research Quarterly*, 11(4):660–685.

CARVER, R. (1985). Is the Degrees of Reading Power Test Valid or Invalid ? *Journal of Reading*, 29(1):34–41.

CARVER, R. (1997). Reading for one second, one minute, or one year from the perspective of rauding theory. *Scientific Studies of Reading*, 1(1):3–43.

CATACH, N. (1985). *Les listes orthographiques de base du français*. Nathan, Paris.

CAYLOR, J., STICHT, T., FOX, L. et FORD, J. (1973). Methodologies for Determining Reading Requirements of Military Occupational Specialties. Rapport technique, Projet n˚73-5, Human Resources Research Organization, Alexandria, VA.

CHALL, J. (1958). Readability : An Appraisal of Research and Application. Rapport technique, Bureau of Educational Research Monographs, Columbus : OH : Ohio State Univ. Press.

CHALL, J. (1988). The Beginning Years. *In* ZAKALUK, B. et SAMUELS, S., éditeurs : *Readability : It's Past, Present, & Future*, pages 2–13. International Reading Association, Newark, Delaware.

CHALL, J., BISSEX, G., CONARD, S. et HARRIS-SHARPLES, S. (1996). *Qualitative assessment of text difficulty : A practical guide for teachers and writers*. Brookline Books, Cambridge, MA.

CHALL, J. et DALE, E. (1995). *Readability Revisited : The New Dale-Chall Readability Formula*. Brookline Books, Cambridge.

CHANIER, T. et SELVA, T. (2000). Génération automatique d'activités lexicales dans le système ALEXIA. *Sciences et Techniques Educatives*, 7(2):385–412.

CHAPELLE, O. et KEERTHI, S. (2008). Multi-class feature selection with support vector machines. Rapport technique, Yahoo Research, Report YR-2008-002.

CHEN, S. et GOODMAN, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13(4):359–393.

CHEN, Y. et LIN, C. (2006). Combining svms with various feature selection strategies. `http://www.csie.ntu.edu.tw/~cjlin/papers/features.pdf`.

CHEVALLIER-WIXLER, D., DUPLEIX, D., JOUETTE, I. et MÈGRE, B. (2007). *Réussir le Dalf, Niveaux C1 et C2*. Didier, Paris.

CHOMSKY, N. (1957). *Syntactic Structures*. Mouton, The Hague.

CHOMSKY, N. (1965). *Aspects of the Theory of Syntax*. MIT Press, Cambridge, Mass.

CICUREL, F. (1991). *Lectures interactives en langues étrangères*. Hachette, Paris.

CLARK, C. (1981). Assessing Comprehensibility : The PHAN System. *The Reading Teacher*, 34(6):670–675.

CLARKE, M. (1979). Reading in Spanish and English : Evidence from adult ESL students. *Language Learning*, 29(1):121–150.

CLAY, M. (1991). *Becoming literate : The construction of inner control*. Heinemann, Portsmouth, NH.

COADY, J. (1979). A psycholinguistic model of the ESL reader. *In* MACKAY, R., BARKMAN, B. et JORDAN, R., éditeurs : *Reading in a second language*. Newbury House, Rowley, MA.

COCCO, C., PITTIER, R., BAVAUD, F. et XANTHOS, A. (2011). Segmentation and Clustering of Textual Sequences : a Typological Approach. *In Proceedings of the International Conference RANLP 2011*.

COKE, E. et ROTHKOPF, E. (1970). Note on a simple algorithm for a computer-produced reading ease score. *Journal of Applied Psychology*, 54(3):208–210.

COLEMAN, E. (1964). The comprehensibility of several grammatical transformations. *Journal of applied psychology*, 48(3):186–190.

COLEMAN, E. et MILLER, G. (1968). A measure of information gained during prose learning. *Reading Research Quarterly*, 3(3):369–386.

COLEMAN, M. et LIAU, T. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

COLLINS-THOMPSON, K. et CALLAN, J. (2004a). A language modeling approach to predicting reading difficulty. *In Proceedings of HLT/NAACL 2004*, pages 193–200, Boston, USA.

COLLINS-THOMPSON, K. et CALLAN, J. (2004b). Information retrieval for language tutoring : An overview of the REAP project. *In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 545–546.

COLLINS-THOMPSON, K. et CALLAN, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56(13):1448–1462.

COLTHEART, M. (1978). Lexical access in simple reading tasks. *In* UNDERWOOD, G., éditeur : *Strategies of information processing*, pages 151–216. Academic Press, London.

COLTHEART, M. (2005). Modeling reading : The dual-route approach. *In* SNOWLING, M. et HULME, C., éditeurs : *The science of reading : A handbook*, pages 6–23. Blackwell Publishing, Oxford.

COLTHEART, M., CURTIS, B., ATKINS, P. et HALLER, M. (1993). Models of reading aloud : Dual-route and parallel-distributed-processing approaches. *Psychological Review*, 100(4):589–608.

COLTHEART, M., DAVELAAR, E., JONASSON, T. et BESNER, D. (1977). Access to the internal lexicon. *In* DORNIC, S., éditeur : *Attention and Performance VI*, pages 535–555. Academic Press, London.

COLTHEART, M., RASTLE, K., PERRY, C., LANGDON, R. et ZIEGLER, J. (2001). DRC : A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, 108(1):204–256.

CONKLIN, K. et SCHMITT, N. (2008). Formulaic sequences : Are they processed more quickly than nonformulaic language by native and nonnative speakers ? *Applied Linguistics*, 29(1):72–89.

CONNINE, C., MULLENNIX, J., SHERNOFF, E. et YELEN, J. (1990). Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology : Learning, Memory, and Cognition*, 16(6):1084–1096.

CONQUET, A. et RICHAUDEAU, F. (1973). Cinq méthodes de mesure de la lisibilité. *Communication et langages*, 17(1):5–16.

CONSEIL DE L'EUROPE (2001). *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*. Hatier, Paris.

CORNAIRE, C. (1985). *La lisibilité : essai d'application de la formule courte d'Henry au français langue étrangère*. Thèse de doctorat, Université de Montréal, Montréal.

CORNAIRE, C. (1988). La lisibilité : essai d'application de la formule courte d'Henry au français langue étrangère. *Canadian Modern Language Review*, 44(2):261–273.

CORNAIRE, C. (1999). *Le point sur la lecture*. CLE international, Paris, réd. édition.

CORNELISSEN, P., TARKIAINEN, A., HELENIUS, P. et SALMELIN, R. (2003). Cortical effects of shifting letter position in letter strings of varying length. *Journal of Cognitive Neuroscience*, 15(5):731–746.

CORSAIN, M., GRANDET, E., MINEMI, E. et RAINOLDI, M. (2007). *Activités pour le CECR, Niveau B2*. CLE International, Paris.

COSTE, D. (1978). Lecture et compétence de communication. *Le Français dans le Monde*, 141:25–35.

COSTE, D. (1990). Un système de dictionnaires électroniques pour les mots simples du français. *Langue Française*, 87:11–22.

COSTE, D., COURTILLON, J., FERENCZI, V., MARTINS-BALTAR, M. et PAPO, E. (1976). *Un Niveau Seuil*. Conseil de l'Europe, Strasbourg.

CRAIN, S. et SHANKWEILER, D. (1988). Syntactic complexity and reading acquisition. *In* DAVIDSON, A. et GREEN, G., éditeurs : *Linguistic complexity and text comprehension : Readability issues reconsidered*, pages 199–221. Erlbaum Associates, Hillsdale, NJ.

CRAWLEY, M. (2007). *The R book*. John Wiley & Sons, Chichester.

CROSSLEY, S., DUFTY, D., MCCARTHY, P. et MCNAMARA, D. (2007). Toward a new readability : A mixed model approach. *In Proceedings of the 29th annual conference of the Cognitive Science Society*, pages 197–202.

CROSSLEY, S., GREENFIELD, J. et MCNAMARA, D. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly*, 42(3):475–493.

CULP, M., JOHNSON, K. et MICHAILIDIS, G. (2006). ada : An r package for stochastic boosting. *Journal of Statistical Software*, 17(2):9.

CZIKO, G. (1980). Language competence and reading strategies : a comparison of first-and second-language oral reading errors. *Language learning*, 30(1):101–114.

DABÈNE, M. (1981). L'acquisition de la compétence de lecture en français, langue étrangère : courants méthodologiques actuels. *System*, 9:215–222.

DAILLE, B. (1995). Combined approach for terminology extraction : lexical statistics and linguistic filtering. Rapport technique, Lancaster University.

DALE, E. (1931). A comparison of two word lists. *Educational Research Bulletin*, 10(18): 484–489.

DALE, E. et CHALL, J. (1948a). A formula for predicting readability. *Educational research bulletin*, 27(1):11–28.

DALE, E. et CHALL, J. (1948b). A formula for predicting readability : Instructions. *Educational research bulletin*, 27(2):37–54.

DALE, E. et CHALL, J. (1949). The concept of readability. *Elementary English*, 26(1):19–26.

DALE, E. et O'ROURKE, J. (1981). *The living word vocabulary : A national vocabulary inventory*. World Book-Childcraft International, Chicago.

DALE, E. et TYLER, R. (1934). A study of the factors influencing the difficulty of reading materials for adults of limited reading ability. *The Library Quarterly*, 4:384–412.

DAMERST, W. (1972). *Clear technical reports*. Harcourt Brace Jovanovich, New York.

DANIELSON, W. et BRYAN, S. (1963). Computer automation of two readability formulas. *Journalism Quarterly*, 40(2):201–205.

DAOUST, F., LAROCHE, L. et OUELLET, L. (1996). SATO-CALIBRAGE : Présentation d'un outil d'assistance au choix et à la rédaction de textes pour l'enseignement. *Revue québécoise de linguistique*, 25(1):205–234.

DAVISON, A. et KANTOR, R. (1982). On the failure of readability formulas to define readable texts : A case study from adaptations. *Reading Research Quarterly*, 17(2):187–209.

DE BEAUGRANDE, R. (1981). Design criteria for process models of reading. *Reading Research Quarterly*, 16(2):261–315.

DE BLESER, R., DUPONT, P., POSTLER, J., BORMANS, G., SPEELMAN, D., MORTELMANS, L. et DEBROCK, M. (2003). The organisation of the bilingual lexicon : a PET study. *Journal of Neurolinguistics*, 16(4-5):439–456.

DE GROOT, A. (1992). Determinants of word translation. *Journal of Experimental Psychology : Learning, Memory, and Cognition*, 18(5):1001–1018.

de LANDSHEERE, G. (1963). Pour une application des tests de lisibilité de Flesch à la langue française. *Le Travail Humain*, 26:141–154.

de LANDSHEERE, G. (1972). Application du cloze test de W.L. Taylor à la langue française. *Scientia Paedagogica Experimentalis*, 9(2):207–256.

de LANDSHEERE, G. (1978). *Le test de closure : mesure de la lisibilité et de la compréhension*. Nathan, Paris.

de LANDSHEERE, G. (1982). *Introduction à la recherche en éducation*. Georges Thone, Liège.

de VILLIERS, P. (1974). Imagery and theme in recall of connected discourse. *Journal of Experimental Psychology*, 103(2):263–268.

DEERWESTER, S., DUMAIS, S., FURNAS, G., LANDAUER, T. et HARSHMAN, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

DEGAND, L. et SANDERS, T. (2002). The impact of relational markers on expository text comprehension in l1 and l2. *Reading and Writing*, 15(7):739–757.

DEGANI, T. et TOKOWICZ, N. (2010). Semantic ambiguity within and across languages : An integrative review. *The Quarterly Journal of Experimental Psychology*, 63(7):1266–1303.

DEHAENE, S. (2007). *Les neurones de la lecture*. Odile Jacob, Paris.

DEHAENE, S., LE CLEC'H, G., POLINE, J., LE BIHAN, D. et COHEN, L. (2002). The visual word form area : a prelexical representation of visual words in the fusiform gyrus. *NeuroReport*, 13(3):321.

DELORY, C. (1989). À propos des indices de lisibilité et des textes de lecture mentale proposés aux enfants en fin d'études primaires. *Scientia Paedagogica Experimentalis*, 26(2):244–256.

DESCHÊNES, A. (1988). *La compréhension et la production de textes*. Presses de l'Université de Québec, Québec.

DESROCHERS, A. et BERGERON, M. (2000). Valeurs de fréquence subjective et d'imagerie pour un échantillon de 1,916 substantifs de la langue française. *Canadian journal of experimental psychology*, 54(4):274–325.

DEVINE, J. (1984). ESL readers' internalized models of the reading process. *In* HANDSCOMBE, J., OREN, R. et TAYLOR, B., éditeurs : *On Tesol '83*, pages 95–108. TESOL, Washington, DC.

DIAS, G., GUILLORÉ, S. et LOPES, J. (2000). Extraction automatique d'associations textuelles à partir de corpora non traités. *In Proceedings of 5th International Conference on the Statistical Analysis of Textual Data*, pages 213–221.

DIJKSTRA, T., GRAINGER, J. et VAN HEUVEN, W. (1999). Recognition of cognates and interlingual homographs : The neglected role of phonology. *Journal of Memory and Language*, 41:496–518.

DIJKSTRA, T., MIWA, K., BRUMMELHUIS, B., SAPPELLI, M. et BAAYEN, H. (2010). How cross-language similarity and task demands affect cognate recognition. *Journal of Memory and Language*, 62(3):284–301.

DIJKSTRA, T. et VAN HEUVEN, W. (2002). The architecture of the bilingual word recognition system : From identification to decision. *Bilingualism : Language and Cognition*, 5(3):175–197.

DOLCH, E. (1928). Vocabulary burden. *The Journal of Educational Research*, pages 170–183.

DOLCH, E. (1939). Vocabulary burden. *Elementary English Review*, 16:135–138.

DOLCH, E. (1948). *Problems in reading*. The Garrard Press, Champaign : Illinois.

DOLLEZ, C. et PONS, C. (2006). *Alter Ego 3*. Hachette, Paris.

DUBAY, W. (2004). *The principles of readability*. Impact Information. Disponible sur http ://www.nald.ca/library/research/readab/readab.pdf.

DUBAY, W. (2007). *Unlocking Language - The Classic Readability Studies*. Information Impact. Disponible sur http ://www.impact-information.com/impactinfo/research/classics.pdf.

DUDA, R., HART, P. et STORK, D. (2001). *Pattern classification*. John Wiley & sons, New York.

DUFFELMEYER, F. (1982). A Comparison of Two Noncomputational Readability Techniques. *The Reading Teacher*, 36(1):4–7.

DUFFY, T. (1985). Readability formulas : What's the use. *Designing usable texts*, 104:113–143.

DUNNING, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74.

EHRLICH, M. et TARDIEU, H. (1986). Le rôle du titre sur le temps de lecture et le rappel de trois types de textes. *Bulletin de Psychologie*, 39(375):397–406.

ELLEY, W. (1969). The assessment of readability by noun frequency counts. *Reading Research Quarterly*, 4(3):411–427.

ELLEY, W. et MANGUBHAI, F. (1983). The impact of reading on second language learning. *Reading Research Quarterly*, 19(1):53–67.

ELLIS, A. et RALPH, M. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems : Insights from connectionist networks. *Journal of Experimental Psychology : Learning, Memory and Cognition*, 26(5):1103–1123.

ESKEY, D. (1998). Holding in the bottom. *In* CARRELL, P., DEVINE, J. et ESKEY, D., éditeurs : *Interactive approaches to second language reading*, pages 93–100. Cambridge Univ. Press, Cambridge.

EVERT, S. et KRENN, B. (2001). Methods for the qualitative evaluation of lexical association measures. *In Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 188–195.

FANG, I. (1966). The "Easy listening formula". *Journal of Broadcasting*, 11(1):63–68.

FANG, I. (1968). By computer flesch's : Reading ease score and a syllable counter. *Behavioral Science*, 13(3):249–251.

FARAWAY, J. (2002). Practical Regression and Anova Using R. *Ann Arbor, MI : self-published (web) URL http ://www.stat.lsa.umich.edu/faraway/book*.

FARR, J., JENKINS, J. et PATERSON, D. (1951). Simplification of Flesch reading ease formula. *Journal of applied psychology*, 35(5):333–337.

FELLBAUM, C. (1998). *WordNet : An electronic lexical database*. MIT press, Cambridge, MA.

FENG, L., ELHADAD, N. et HUENERFAUTH, M. (2009). Cognitively motivated features for readability assessment. *In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 229–237.

FENG, L., JANSCHE, M., HUENERFAUTH, M. et ELHADAD, N. (2010). A Comparison of Features for Automatic Readability Assessment. *In COLING 2010 : Poster Volume*, pages 276–284.

FERRAND, L. (1996). The masked repetition priming effect dissipates when increasing the inter-stimulus interval : Evidence from word naming. *Acta Psychologica*, 91(1):15–25.

FERRAND, L. (2007). *Psychologie cognitive de la lecture*. De Boeck, Bruxelles.

FERRETTI, T., KUTAS, M. et MCRAE, K. (2007). Verb aspect and the activation of event knowledge. *Journal of Experimental Psychology : Learning, Memory, and Cognition*, 33(1):182–196.

FIGUERAS, N. (2007). The cefr, a lever for the improvement of language professionals in europe. *Modern Language Journal*, 91(4):673–675.

FISET, D. et ARGUIN, M. (2002). L'approche connexioniste en neuropsychologie cognitive : le cas de la lecture et des dyslexies acquises. *Revue Québécoise de Psychologie*, 23(2): 181–205.

FLESCH, R. (1946). *The Art of Plain Talk*. Harper & Brothers, New York.

FLESCH, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

FLESCH, R. (1949). New Facts about Readability. *College English*, 10(4):225–226.

FLESCH, R. (1950). Measuring the level of abstraction. *Journal of Applied Psychology*, 34(6):384–390.

FLESCH, R. (1951). Reply to criticism by jenkins and jones. *Journal of Applied Psychology*, 35(1):69.

FLUMIAN, C., LABASCOULE, J. et ROYER, C. (2004). *Rond Point 2*. Difusión, Barcelone.

FOLTZ, P., KINTSCH, W. et LANDAUER, T. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2):285–307.

FONT, N. et LAVAUR, J. (2004). Effets de la fréquence du voisinage orthographique interlangue lors de la reconnaissance visuelle de mots chez les bilingues. *L'année psychologique*, 104(3):377–405.

FORBES, F. et COTTLE, W. (1953). A new method for determining readability of standardized tests. *Journal of Applied Psychology*, 37(3):185–190.

FORSTER, K. et DAVIS, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology : Learning, Memory, and Cognition*, 10(4): 680–698.

FOUCAMBERT, J. (1983). Une lecture de la lisibilité. *les Actes de lecture*, 3:37–39.

FOUCART, T. (2006). Colinéarité et régression linéaire. *Mathématiques et sciences humaines*, (173):5–25.

FOUNTAS, I. et PINNELL, G. (1999). *Matching books to readers : Using leveled books in guided reading*. Heinemann, Portsmouth.

FRANÇOIS, T. (2009a). Combining a statistical language model with logistic regression to predict the lexical and syntactic difficulty of texts for FFL. *In Proceedings of the 12th Conference of the EACL : Student Research Workshop*, pages 19–27.

FRANÇOIS, T. (2009b). Modèles statistiques pour l'estimation automatique de la difficulté de textes de FLE. *In 11eme Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*.

FRANÇOIS, T. (2011). La lisibilité computationnelle : un renouveau pour la lisibilité du français langue première et seconde ? *International Journal of Applied Linguistics (ITL)*, 160.

FRANÇOIS, T. et WATRIN, P. (2011a). On the contribution of MWE-based features to a readability formula for French as a foreign language. *In Proceedings of the International Conference RANLP 2011*.

FRANÇOIS, T. et WATRIN, P. (2011b). Quel apport des unités polylexicales dans une formule de lisibilité pour le français langue étrangère ? *In Actes de la 18e Conférence sur le Traitement automatique des Langues Naturelles (TALN 2011)*.

FRASE, L. (1969). Paragraph organization of written materials : The influence of conceptual clustering upon the level and organization of recall. *Journal of Educational Psychology*, 60(5):394–401.

FREROT, J.-L. et GIRARDET, J. (1997). *Panomara 3*. CLE International, Paris.

FREUND, Y. et SCHAPIRE, R. (1996). Experiments with a new boosting algorithm. *In Machine Learning : Proceedings of the Thirteenth International Conference*, pages 148–156.

FRUNZA, O. et INKPEN, D. (2007). A tool for detecting French-English cognates and false friends. *In Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles*, pages 91–100.

FRY, E. (1963). *Teaching faster reading*. Cambridge University Press, London.

FRY, E. (1968). A readability formula that saves time. *Journal of reading*, 11(7):513–578.

FRY, E. (1977). Fry's readability graph : Clarifications, validity, and extension to level 17. *Journal of Reading*, 21(3):242–252.

FRY, E. (1989). Reading formulas : maligned but valid. *Journal of reading*, 32(4):292–297.

FRY, E. (1990). A readability formula for short passages. *Journal of Reading*, 33(8):594–597.

FRY, E. (2002). Readability versus Leveling : Both of These Procedures Can Help Teachers Select Books for Readers at Different Stages. *The Reading Teacher*, 56(3):286–292.

GAIES, S. (1979). Linguistic imput in formal second language learning : the issues of syntactic gradation and readability in ESL materials. *Tesol Quarterly*, 13(1):41–50.

GALE, W. et SAMPSON, G. (1995). Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, 2(3):217–237.

GALLEY, M. et MCKEOWN, K. (2003). Improving word sense disambiguation in lexical chaining. *In International Joint Conference on Artificial Intelligence*, volume 18, pages 1486–1488.

GAONAC'H, D. (2005). La lecture en langue étrangère : un tour d'horizon d'une problématique de psychologie cognitive. *Acquisition et Interaction en Langue Étrangère [En ligne]*.

GÉLINAS CHEBAT, C., PRÉFONTAINE, C., LECAVALLIER, J. et CHEBAT, J.-C. (1993). Lisibilité - intelligibilité de documents d'information. *Cahier de recherche*, 3:19–35.

GERHAND, S. et BARRY, C. (1998). Word frequency effects in oral reading are not merely age-of-acquisition effects in disguise. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 24(2):267–283.

GERHAND, S. et BARRY, C. (1999). Age of acquisition, word frequency, and the role of phonology in the lexical decision task. *Memory & Cognition*, 27(4):592–602.

GERNSBACHER, M. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology : General*, 113(2):256–281.

GIASSON, J. (1990). *La compréhension en lecture*. De Boeck, Bruxelles.

GIASSON, J. (1997). *La lecture. De la théorie à la pratique*. De Boeck, Bruxelles.

GILLIE, P. (1957). A simplified formula for measuring abstraction in writing. *Journal of applied psychology*, 41(4):214–217.

GILLILAND, J. (1972). *Readability*. Hodder and Stoughton, London.

GIRARD, D. (1995). *Enseigner les langues : méthodes et pratiques*. Bordas, Paris.

GIRARDET, J. et CRIDLIG, J.-M. (1996a). *Panorama 1*. CLE International, Paris.

GIRARDET, J. et CRIDLIG, J.-M. (1996b). *Panorama 2*. CLE International, Paris.

GOODMAN, K. (1967). Reading : A psycholinguistic guessing game. *Journal of the Reading Specialist*, 6:126–35.

GOODMAN, K. (1968). The psycholinguistic nature of the reading process. *In* GOODMAN, K., éditeur : *The psycholinguistic nature of the reading process*. Wayne State University Press, Detroit.

GOODMAN, K. (1969). Analysis of oral reading miscues : Applied psycholinguistics. *Reading Research Quarterly*, pages 9–30.

GORMAN, A. (1961). Recognition memory for nouns as a function of abstractness and frequency. *Journal of Experimental Psychology*, 61(1):23–29.

GOSWAMI, U. (2004). Neuroscience and education. *British Journal of Educational Psychology*, 74:1–14.

GOUGENHEIM, G., MICHÉA, R., RIVENC, P. et SAUVAGEOT, A. (1964). *L'élaboration du français fondamental (1er degré)*. Didier, Paris.

GOUGH, P. (1965). Grammatical transformations and speed of understanding. *Journal of verbal learning and verbal behavior*, 4(2):107–111.

GOUGH, P. (1972). One Second of Reading. *In* KAVANAGH, J. F. et MATTINGLY, I. G., éditeurs : *Language by ear and by eye*, pages 331–358. MIT Press, Cambridge.

GOUGH, P. (1984). Word recognition. *In* PEARSON, P., éditeur : *Handbook of reading research*, pages 225–253. Longman, New York.

GRABE, W. (1991). Current developments in second language reading research. *Tesol Quarterly*, pages 375–406.

GRABE, W. (1998). Reassessing the term "interactive". *In* CARRELL, P., DEVINE, J. et ESKEY, D., éditeurs : *Interactive Approaches to Second Language Reading*, pages 56–70. Cambridge University Press, Cambridge.

GRABE, W. (2005). Reading in a second language. *In* KAPLAN, R., GRABE, W., SWAIN, M. et TUCKER, G., éditeurs : *The Oxford Handbook of Applied Linguistics*, pages 49–59. Oxford University Press, Oxford.

GRAESSER, A., MCNAMARA, D., LOUWERSE, M. et CAI, Z. (2004). Coh-Metrix : Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202.

GRANOWSKY, A. et BOTEL, M. (1974). Background for a new syntactic complexity formula. *The Reading Teacher*, 28(1):31–35.

GRAY, W. et LEARY, B. (1935). *What makes a book readable*. University of Chicago Press, Chicago : Illinois.

GREENFIELD, J. (2004). Readability formulas for EFL. *Japan Association for Language Teaching*, 26(1):5–24.

GROMER, B. et WEISS, M. (1990). *Lire*. Paris, Armand Collin.

GUBERINA, P. (1984). Bases théoriques de la méthode audiovisuelle structuro-globale (méthode Saint-Cloud-Zagreb). Une linguistique de la parole. *In* COSTE, D., éditeur : *Aspects d'une politique de diffusion du français langue étrangère depuis 1945. Matériaux pour une histoire*, pages 85–98. Hatier, Paris.

GUILFORD, J. (1965). *Fundamental statistics in psychology and education*. McGraw-Hill, New-York.

GUNNING, R. (1952). *The technique of clear writing*. McGraw-Hill, New York.

GUTHRIE, J. (1973). Reading comprehension and syntactic responses in good and poor readers. *Journal of Educational Psychology*, 65(3):294–299.

GUYON, I., WESTON, I., WESTON, J., BARNHILL, S. et VAPNIK, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422.

HAHNE, A. (2001). What's different in second-language processing ? Evidence from event-related brain potentials. *Journal of Psycholinguistic Research*, 30(3):251–266.

HAHNE, A. et FRIEDERICI, A. (1999). Electrophysiological evidence for two steps in syntactic analysis : Early automatic and late controlled processes. *Journal of Cognitive Neuroscience*, 11(2):194–205.

HALLIDAY, M. et HASAN, R. (1976). *Cohesion in English*. Longman Group Ltd, London.

HARM, M. et SEIDENBERG, M. (2004). Computing the meanings of words in reading : Cooperative division of labor between visual and phonological processes. *Psychological Review*, 111(3):662–720.

HARRIS, A. et JACOBSON, M. (1972). *Basic Elementary Reading Vocabularies*. MacMillan Publishing, New York.

HARRIS, A. et JACOBSON, M. (1973). The Harris-Jacobson primary readability formulas. *In Proceedings of the annual meeting of the College Reading Association*, Denver, Colorado.

HARRIS, A. et JACOBSON, M. (1974). Revised Harris-Jacobson readability formulas. *In Proceedings of the annual meeting of the College Reading Association*, Bethesda, Maryland.

HARRIS, A. et JACOBSON, M. (1976). Predicting twelfth graders' comprehension scores. *Journal of Reading*, 20(1):43–46.

HARRIS, A. et JACOBSON, M. (1979). A framework for readability research : Moving beyond Herbert Spencer. *Journal of Reading*, 22(5):390–398.

HARRISSON, C. (1980). *Readability in the classroom*. Cambridge University Press, Cambridge.

HASTIE, T., TIBSHIRANI, R. et FRIEDMAN, J. (2001). *The elements of statistical learning : data mining, inference and prediction*. Springer-Verlag, New York.

HEILMAN, M., COLLINS-THOMPSON, K., CALLAN, J. et ESKENAZI, M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. *In Proceedings of NAACL HLT*, pages 460–467.

HEILMAN, M., COLLINS-THOMPSON, K. et ESKENAZI, M. (2008). An analysis of statistical models and features for reading difficulty prediction. *In Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–8.

HENRY, G. (1969). Étude de la composante vocabulaire dans les formules de lisibilité. *Scientia Paedagogica Experimentalis*, 5(2):221–228.

HENRY, G. (1975). *Comment mesurer la lisibilité*. Labor, Bruxelles.

HENRY, G. (1979). Une méthode de mesure par ordinateur de la lisibilité des textes français. *Scientia Paedagogica Experimentalis*, 16(1):52–58.

HENRY, G. (1980). Lisibilité et compréhension. *Communication et langages*, 45(2):7–15.

HINO, Y. et LUPKER, S. (1996). Effects of polysemy in lexical decision and naming : An alternative to lexical access accounts. *Journal of Experimental Psychology : Human Perception and Performance*, 22(6):1331–1356.

HINO, Y., LUPKER, S. et PEXMAN, P. (2002). Ambiguity and synonymy effects in lexical decision, naming, and semantic categorization tasks : Interactions between orthography, phonology, and semantics. *Journal of Experimental Psychology : Learning, Memory, and Cognition*, 28(4):686–713.

HINTON, G. (1989). Connectionist learning procedures. *Artificial intelligence*, 40:185–234.

HOOVER, W. et GOUGH, P. (1990). The simple view of reading. *Reading and Writing : An Interdisciplinary Journal*, 2:127–160.

HOOVER, W. et TUNMER, W. (1993). The Components of Reading. *In* THOMPSON, G., TUNMER, W. et NICHOLSON, T., éditeurs : *Reading acquisition processes*, pages 1–19. Multilingual Matters, Clevedon.

HORELLOU-LAFARGE, C. et SEGRÉ, M. (2003). *Sociologie de la lecture*. La Découverte, Paris.

HOSENFELD, C. (1984). Case studies of ninth grade readers. *In* ALDERSON, J. et URQUHART, A., éditeurs : *Reading in a Foreign Language*, pages 231–249. Longman, New York.

HOSMER, D. et LEMESHOW, S. (1989). *Applied Logistic Regression*. Wiley, New York.

HOWELL, D. (2008). *Méthodes statistiques en sciences humaines, 6ème édition*. De Boeck, Bruxelles.

HOWES, D. et SOLOMON, R. (1951). Visual duration threshold as a function of word probability. *Journal of Experimental Psychology*, 41(40):1–4.

HSU, C., CHANG, C. et LIN, C. (2009). A practical guide to support vector classification. Rapport technique, Department of Computer Science and Information Engineering, National Taiwan University.

HUEY, E. (1908). *The psychology and pedagogy of reading*. Macmillan, New York.

HUNT, K. (1965). *Grammatical Structures Written at Three Grade Levels*. Rapport technique, National Council of Teachers of English. Research Report No. 3, Champaign, IL.

HYMES, D. H. (1991). *Vers la compétence de communication*. Hatier/Credif, Paris.

IDE, N. (1998). The xml framework and its implications for the development of natural language processing tools. *In Proceedings of the COLING Workshop on Using Toolsets and Architectures to Build NLP Systems*.

IDE, N., BONHOMME, P. et ROMARY, L. (2000). Xces : An xml-based encoding standard for linguistic corpora. *In Proceedings of the Second International Conference on Language Resources and Evaluation*, pages 825–830.

IDE, N. et VERONIS, J. (1995). *Text encoding initiative : background and contexts*. Kluwer Academic Publishers, Dordrecht.

INHOFF, A. et RAYNER, K. (1986). Parafoveal word processing during eye fixations in reading : Effects of word frequency. *Perception & Psychophysics*, 40:431–439.

IRWIN, D. (1998). Lexical processing during saccadic eye movements. *Cognitive Psychology*, 36(1):1–27.

IRWIN, J. et DAVIS, C. (1980). Assessing readability : The checklist approach. *Journal of Reading*, 24(2):124–130.

ISEL, F. (2007). Syntactic and referential processes in second-language learners : event-related brain potential evidence. *NeuroReport*, 18(18):1885–89.

IWAHORI, Y. (2008). Developing reading fluency : A study of extensive reading in efl. *Reading in a Foreign Language*, 20(1):70–91.

JACOBSON, M. (1965). Reading difficulty of physics and chemistry textbooks. *Educational and Psychological Measurement*, 25(2):449–457.

JACOBSON, M. (1974). Predicting reading difficulty from spelling. *Spelling Progress Bulletin*, 14:8–10.

JACOBSON, M., KIRKLAND, C. et SELDEN, R. (1978a). An examination of the McCall-Crabbs standard test lessons in reading. *Journal of Reading*, 22(3):224–230.

JACOBSON, M., KIRKLAND, C. et SELDEN, R. (1978b). An examination of the McCall-Crabbs standard test lessons in reading. *Journal of Reading*, 22(3):224–230.

JENKINS, J. et JONES, R. (1951). Flesch's "measuring the level of abstraction.". *Journal of Applied Psychology*, 35(1):68.

JIANG, N. et NEKRASOVA, T. (2007). The processing of formulaic sequences by second language speakers. *The Modern Language Journal*, 91(3):433–445.

JOBARD, G., CRIVELLO, F. et TZOURIO-MAZOYER, N. (2003). Evaluation of the dual route theory of reading : a metanalysis of 35 neuroimaging studies. *Neuroimage*, 20(2):693–712.

JOHNSON, A.-M., MENAND, R., BERTHET, A. et KIZIRIAN, V. (2004). *Taxi ! 3*. Hachette, Paris.

JOHNSON, C. (1927). Vocabulary Difficulty and Textbook Selection. *Modern Language Journal*, 11(5):290–297.

JOHNSON, G. (1930). An objective method of determining reading difficulty. *The Journal of Educational Research*, pages 283–287.

JOHNSON, J. et NEWPORT, E. (1989). Critical period effects in second language learning : The influence of maturational state on the acquisition of English as a second language. *Cognitive psychology*, 21(1):60–99.

JOHNSON, R. et BOND, G. (1950). Reading ease of commonly used tests. *Journal of Applied Psychology*, 34(5):319–324.

JONGSMA, E. (1969). *The cloze procedure : a survey of the research*. Rapport technique, Indiana University, Bloomington. School of Education.

JOUVE, V. (1993). *La lecture*. Hachette, Paris.

JURAFSKY, D. et MARTIN, J. (2009). *Speech and language processing*. Prentice Hall, Upper Saddle River.

JUST, M. et CARPENTER, P. (1980). A theory of reading : From eye fixations to comprehension. *Psychological review*, 87(4):329–354.

KAAN, E., HARRIS, A., GIBSON, E. et HOLCOMB, P. (2000). The P600 as an index of syntactic integration difficulty. *Language and cognitive processes*, 15(2):159–201.

KAMMANN, R. et STREETER, L. (1971). Two meanings of word abstractness. *Journal of Verbal Learning and Verbal Behavior*, 10(3):303–306.

KANDEL, L. et MOLES, A. (1958). Application de l'indice de Flesch à la langue française. *Cahiers Études de Radio-Télévision*, 19:253–274.

KANUNGO, T. et ORR, D. (2009). Predicting the readability of short web summaries. *In Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 202–211.

KATE, R., LUO, X., PATWARDHAN, S., FRANZ, M., FLORIAN, R., MOONEY, R., ROUKOS, S. et WELTY, C. (2010). Learning to predict readability using diverse linguistic features. *In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 546–554.

KEENAN, J., BAILLET, S. et BROWN, P. (1984). The effects of causal cohesion on comprehension and memory. *Journal of Verbal Learning and Verbal Behavior*, 23(2):115–126.

KEMPER, S. (1983). Measuring the inference load of a text. *Journal of Educational Psychology*, 75(3):391–401.

KEMPER, S., JACKSON, J., CHEUNG, H. et ANAGNOPOULOS, C. (1993). Enhancing older adults' reading comprehension. *Discourse Processes*, 16(4):405–428.

KIBBY, M. (1981). Test Review : The Degrees of Reading Power. *Journal of Reading*, 24(5):416–427.

KIDWELL, P., LEBANON, G. et COLLINS-THOMPSON, K. (2009). Statistical estimation of word acquisition with application to readability prediction. *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, volume 2, pages 900–909.

KILGARRIFF, A. (2005). Language is never ever ever random. *Corpus linguistics and linguistic theory*, 1(2):263–276.

KINCAID, J., FISHBURNE, R., RODGERS, R. et CHISSOM, B. (1975). *Derivation of new readability formulas for navy enlisted personnel*. Rapport technique, n˚8-75, Research Branch Report.

KINTCH, W. (1998). *Comprehension : A paradigm for cognition*. Cambridge, Cambridge University Press.

KINTSCH, W. (1979). On modeling comprehension. *Educational Psychologist*, 14(1):3–14.

KINTSCH, W. (1988). The role of knowledge in discourse comprehension : A construction-integration model. *Psychological review*, 95(2):163–182.

KINTSCH, W. et KEENAN, J. (1973). Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive psychology*, 5(3):257–274.

KINTSCH, W., KOZMINSKY, E., STREBY, W., MCKOON, G. et KEENAN, J. (1975). Comprehension and recall of text as a function of content variables1. *Journal of Verbal Learning and Verbal Behavior*, 14(2):196–214.

KINTSCH, W., MANDEL, T. et KOZMINSKY, E. (1977). Summarizing scrambled stories. *Memory & Cognition*, 5:547–552.

KINTSCH, W. et MILLER, J. (1984). Readability : A view from cognitive psychology. *In* FLOOD, J., éditeur : *Understanding reading comprehension*, pages 220–232. International Reading Association, Newark.

KINTSCH, W. et van DIJK, T. (1978). Toward a model of text comprehension and production. *Psychological review*, 85(5):363–394.

KINTSCH, W. et VIPOND, D. (1979). Reading comprehension and readability in educational practice and psychological theory. *In* NILSSON, L., éditeur : *Perspectives on Memory Research*, pages 329–365. Lawrence Erlbaum, Hillsdale, NJ.

KIRSCH, I. et MOSENTHAL, P. (1990). Exploring document literacy : Variables underlying the performance of young adults. *Reading Research Quarterly*, 25(1):5–30.

KISTER, L. et JACQUEY, E. (2007). NdeN et acquisition d'informations lexicales à partir du Trésor de la Langue Française Informatisé. *Cognition, Représentation, Langage*, 5(1).

KLARE, G. (1952). Measures of the readability of written communication : an evaluation. *Journal of Educational Psychology*, 43(7):385–399.

KLARE, G. (1963). *The Measurement of Readability*. Iowa State University Press, Ames, IA.

KLARE, G. (1968). The role of word frequency in readability. *Elementary English*, pages 12–22.

KLARE, G. (1974). Assessing readability. *Reading Research Quarterly*, 10(1):62–102.

KLARE, G. (1984). Readability. *In* PEARSON, P., BARR, R., KAMIL, M. L., MOSENTHAL, P. et DYKSTRA, R., éditeurs : *Handbook of Reading Research*, pages 681–744. Longman, New York.

KLARE, G. (1988). The Formative Years. *In* ZAKALUK, B. et SAMUELS, S., éditeurs : *Readability : It's Past, Present, & Future*, pages 14–34. International Reading Association, Newark, Delaware.

KLARE, G., ROWE, P., ST. JOHN, M. et STOLUROW, L. (1969). Automation of the Flesch reading ease readability formula, with various options. *Reading research quarterly*, 4(4): 550–559.

KLARE, G., SINAIKO, H. et STOLUROW, L. (1972). The cloze procedure : A convenient readability test for training materials and translations. *Applied Psychology*, 21(2):77–105.

KOBER-KLEINERT, C., PARIZET, M.-L. et POISSON-QUINTON, S. (2007). *Activités pour le CECR niveau C1/C2*. CLE International, Paris.

KODA, K. (2005). *Insights into second language reading : A cross-linguistic approach*. Cambridge University Press, Cambridge.

KOLERS, P. (1970). Three Stages of Reading. *In* LEVIN, H. et WILLIAMS, J., éditeurs : *Basic Studies on Reading*, pages 90–118. Basic Book, New York.

KONDRAK, G. (2004). Combining evidence in cognate identification. *In Proceedings of Canadian AI 2004 : 17th Conference of the Canadian Society for Computational Studies of Intelligence*, pages 44–59.

KOUNIOS, J. et HOLCOMB, P. (1994). Concreteness effects in semantic processing : ERP evidence supporting dual-coding theory. *Journal of Experimental Psychology : Learning, Memory, and Cognition*, 20(4):804–823.

KOZLOFF, M. (2002). Rhetoric and revolution : Kenneth Goodman's "Psycholinguistic guessing game". *Direct Instruction News*, 2(2):34–41.

KRASHEN, S. (1989). We acquire vocabulary and spelling by reading : Additional evidence for the input hypothesis. *The Modern Language Journal*, 73(4):440–464.

KUČERA, H. et FRANCIS, W. (1967). *Computational analysis of present-day American English*. Brown University Press, Providence.

KUTAS, M. et HILLYARD, S. (1980). Reading senseless sentences : Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205.

LABASCOULE, J., LAUSE, C. et ROYER, C. (2005). *Rond-Point 1 (3e éd.)*. Difusión, Barcelone.

LABASSE, B. (1999). La lisibilité rédactionnelle : fondements et perspectives. *Communication et langages*, 121(1):86–103.

LABERGE, D. et SAMUELS, S. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6(2):293–323.

LANDAUER, T., FOLTZ, P. et LAHAM, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2):259–284.

LAROCHE, J. (1979). Readability measurement for foreign-language materials. *System*, 7(2):131–135.

LARSON, K. (2004). The Science of Word Recognition : Or How I Learned to Stop Worrying and Love the Bouma. *Microsoft Typography site. Accès : Août 2009*.

LAUFER, B. (1997). What's in a word that makes it hard or easy : Some intralexical factors that affect the learning of words. *In* SCHMITT, N. et MCCARTHY, M., éditeurs : *Vocabulary : Description, Acquisition and Pedagogy*, pages 140–155. Cambridge University Press, Cambridge.

LEBARBIER, E. et MARY-HUARD, T. (2004). Le critère bic : fondements théoriques et interprétation. Rapport technique, INRIA, Institut national de recherche en informatique et automatique.

LEE, H., GAMBETTE, P., MAILLÉ, E. et THUILLIER, C. (2010). Densidées : calcul automatique de la densité des idées dans un corpus oral. *In Actes de la douxième Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des langues (RECITAL)*.

LEHMANN, D. (1994). Lire en français langue étrangère. *In* COSTE, D., éditeur : *Vingt ans dans l'évolution de la didactique des langues (1968-1988)*, pages 84–99. Hatier-Didier / Credif, Paris.

LEMHÖFER, K., DIJKSTRA, T., SCHRIEFERS, H., BAAYEN, R., GRAINGER, J. et ZWITSERLOOD, P. (2008). Native language influences on word recognition in a second language : A megastudy. *Journal of Experimental Psychology. Learning, Memory and Cognition*, 34(1):12–31.

LESCURE, R., CHENARD, S., MUBANGA BEYA, A., RAUSCH, A. et VEY, P. (2007). *DALF C1/C2, 250 activités - Le Nouvel Entraînez-vous*. CLE International, Paris.

LEVIN, H. et KAPLAN, E. (1970). Grammatical structure and reading. *In* LEVIN, H. et WILLIAMS, J., éditeurs : *Basic Studies on Reading*, pages 119–133. Basic Book, New York.

LÉVY-SCHOEN, A. et O'REGAN, K. (1978). Les mouvements des yeux au cours de la lecture. *L'année psychologique*, 78(2):459–491.

LEWERENZ, A. (1935). A vocabulary grade placement formula. *The Journal of Experimental Educational*, 3(3):236.

LEWIS-BECK, M. (1993). *Experimental Design and Methods*, volume 3 de *International Handbooks of Quantitative Applications in the Social Sciences*. Sage Publications, Singapore.

LIAU, T., BASSIN, C., MARTIN, C. et COLEMAN, E. (1976). Modification of the Coleman readability formulas. *Journal of Literacy Research*, 8(4):381–386.

LIN, D. (1999). Automatic identification of non-compositional phrases. *In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 317–324.

LITTLE, D. (2006). The common european framework of reference for languages : Content, purpose, origin, reception and impact. *Language Teaching*, 39(3):167–190.

LITTLE, D. (2007). The common european framework of reference for languages : Perspectives on the making of supranational language education policy. *The Modern Language Journal*, 91(4):645–655.

LIVELY, B. et PRESSEY, S. (1923). A method for measuring the “vocabulary burden” of textbooks. *Educational Administration and Supervision*, 9:389–398.

LOISEAU, Y. et MÉRIEUX, R. (2004). *Connexions 1*. Didier, Paris.

LORGE, I. (1939). Predicting reading difficulty of selections for children. *Elementary English Review*, 16(6):229–33.

LORGE, I. (1944a). Predicting readability. *the Teachers College Record*, 45(6):404–419.

LORGE, I. (1944b). Word lists as background for communication. *The Teachers College Record*, 45(8):543–552.

LORGE, I. (1948). The Lorge and Flesch readability formulae : a correction. *School and Society*, 67:141–142.

LOWE, A. (1979). ThUS : A Non-Formula Readability Measure. *Reading Improvement*, 16(2):155–57.

LUPKER, S. (2005). Visual word recognition : Theories and findings. *In* SNOWLING, M. et HULME, C., éditeurs : *The science of reading : A handbook*, pages 39–60. Blackwell Publishing, Oxford.

MACDONALD, N., FRASE, L., GINGRICH, P. et KEENAN, S. (1982). The writer’s workbench : Computer aids for text analysis. *Educational psychologist*, 17(3):172–179.

MACGINITIE, W. et TRETIAK, R. (1971). Sentence depth measures as predictors of reading difficulty. *Reading research quarterly*, 6:364–377.

MACKAY, R., BARKMAN, B. et JORDAN, R., éditeurs (1979). *Reading in a second language : hypotheses, organization and practice*. Newbury House, Rowley, MA.

MACWHINNEY, B. (1997). Second language acquisition and the competition model. *In* de GROOT, A. et KROLL, J., éditeurs : *Tutorials in bilingualism : Psycholinguistic perspectives*, pages 113–142. Lawrence Erlbaum, Mahwah, NJ.

MANNING, C., RAGHAVAN, P. et SCHÜTZE, H. (2008). *Introduction to information retrieval*. Cambridge University Press, Cambridge, UK.

MARINKOVIC, K., DHOND, R., DALE, A., GLESSNER, M., CARR, V. et HALGREN, E. (2003). Spatiotemporal dynamics of modality-specific and supramodal word processing. *Neuron*, 38(3):487–497.

MARSHALL, J. et NEWCOMBE, F. (1973). Patterns of paralexia : A psycholinguistic approach. *Journal of Psycholinguistic Research*, 2(3):175–199.

MARSHALL, N. (1979). Readability and comprehensibility. *Journal of Reading*, 22(6):542–544.

MASON, R. et JUST, M. (2004). How the brain processes causal inferences in text. *Psychological Science*, 15(1):1–7.

MASTROPIERI, M., LEINART, A. et SCRUGGS, T. (1999). Strategies to increase reading fluency. *Intervention in School and Clinic*, 34(5):278–283.

MATIN, E. (1974). Saccadic suppression : A review and an analysis. *Psychological Bulletin*, 81(12):899–917.

MAXWELL, M. (1978). Readability : Have we gone too far ? *Journal of reading*, 21(6):525–530.

MCCALLUM, D. et PETERSON, J. (1982). Computer-based readability indexes. *In Proceedings of the ACM'82 conference*, pages 44–48.

MCCLELLAND, J. (1979). On the time relations of mental processes : An examination of systems of processes in cascade. *Psychological Review*, 86(4):287–330.

MCCLELLAND, J. et RUMELHART, D. (1981). An interactive activation model of context effects in letter perception : Part 1. An account of basic findings. *Psychological Review*, 88(5):375–407.

MCCLUSKY, H. (1934). A Quantitative Analysis of the Difficulty of Reading Materials. *The Journal of Educational Research*, 28:276–282.

MCCONKIE, G. et RAYNER, K. (1975). The span of the effective stimulus during a fixation in reading. *Perception and Psychophysics*, 17(6):578–586.

MCCONKIE, G. et RAYNER, K. (1976). Asymmetry of the perceptual span in reading. *Bulletin of the Psychonomic Society*, 8(5):365–368.

MCCULLAGH, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):109–142.

MCLAUGHLIN, G. (1969). SMOG grading : A new readability formula. *Journal of reading*, 12(8):639–646.

MCNAMARA, D., LOUWERSE, M., MCCARTHY, P. et GRAESSER, A. (2010). Coh-Metrix : Capturing linguistic features of cohesion. *Discourse Processes*, 47(4):292–330.

MCNEMAR, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

MEIR, R. et RÄTSCH, G. (2003). An introduction to boosting and leveraging. *In* MENDELSON, S. et SMOLA, A., éditeurs : *Advanced lectures on machine learning*, numéro 2600, pages 118–183. Springer-Verlag, Berlin.

MENAND, R., BERTHET, A. et KIZIRIAN, V. (2003). *Taxi ! 2*. Hachette, Paris.

MESNAGER, J. (1986a). La lisibilité dans la littérature enfantine. *Les actes de lecture*, 13:31–39.

MESNAGER, J. (1986b). La lisibilité dans la littérature enfantine. *Les actes de lecture*, 14:25–33.

MESNAGER, J. (1989). Lisibilité des textes pour enfants : un nouvel outil ? *Communication et Langages*, 79:18–38.

MESNAGER, J. (2002). Pour une étude de la difficulté des textes : la lisibilité revisitée. *Le Français aujourd'hui*, 137:29–42.

MEYER, B. (1982). Reading research and the composition teacher : The importance of plans. *College composition and communication*, 33(1):37–49.

MEYER, B., BRANDT, D. et BLUTH, G. (1980). Use of author's textual schema : Key for ninth graders' comprehension. *Reading Research Quarterly*, 16:72–103.

MEYER, B., MARSISKE, M. et WILLIS, S. (1993). Text processing variables predict the readability of everyday documents read by older adults. *Reading research quarterly*, 28(3):235.

MEYER, D. et SCHVANEVELDT, R. (1971). Facilitation in recognizing pairs of words : Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90(2):227–234.

MICHEL, J., SHEN, Y., AIDEN, A., VERES, A., GRAY, M., TEAM, T. G. B., PICKETT, J., HOIBERG, D., CLANCY, D., NORVIG, P., ORWANT, J., PINKER, S., NOWAK, M. et AIDEN, E. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

MILLER, G. (1956). The magical number seven plus or minus two : some limits on our capacity for processing information. *Psychological review*, 63(2):81–97.

MILLER, G. et CHOMSKY, N. (1963). Finitary Models of Language Users. *In* LUCE, R., BUSH, R. et GALANTER, E., éditeurs : *Handbook of mathematical psychology*, pages 419–492. Wiley, New York.

MILLER, G. et COLEMAN, E. (1967). A set of thirty-six prose passages calibrated for complexity. *Journal of Verbal Learning and Verbal Behavior*, 6(6):851–854.

MILLER, J. et KINTSCH, W. (1980). Readability and recall of short prose passages : A theoretical analysis. *Journal of Experimental Psychology : Human Learning and Memory*, 6(4):335–354.

MILLIS, M. et BUTTON, S. (1989). The effect of polysemy on lexical decision time : Now you see it, now you don't. *Memory & Cognition*, 17(2):141–147.

MILTSAKAKI, E. (2009). Matching readers' preferences and reading skills with appropriate web texts. *In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics : Demonstrations Session*, pages 49–52.

MITKOV, R., PEKAR, V., BLAGOEV, D. et MULLONI, A. (2007). Methods for extracting and classifying pairs of cognates and false friends. *Machine translation*, 21(1):29–53.

MOIRAND, S. (1976). Approche globale de textes écrits. *Études de linguistique appliquée*, 23:88–105.

MOIRAND, S. (1979). *Situations d'écrit : compréhension/production en français langue étrangère*. CLE International, Paris.

MONSELL, S. (1991). The nature and locus of word frequency effects in reading. *In* BESNER, D. et HUMPHREYS, G., éditeurs : *Basic processes in reading : Visual word recognition*, pages 148–197. Lawrence Erlbaum Associates Inc., Hillsdale, NJ.

MORGAN, J. L. et SELLNER, M. (1980). Discourse and linguistic theory. *In* SPIRO, R., BERTRAM, B. et BREWER, W., éditeurs : *Theoretical issues in reading comprehension*, pages 165–200. Lawrence Erlbaum Associates, Hillsdale, NJ.

MORRIS, J. (1972). *Make yourself clear*. McGraw-Hill, New York.

MORRISON, C. et ELLIS, A. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 21(1):116–133.

MORRISON, D. (1968). PATRICIA - practical algorithm to retrieve information coded in alphanumeric. *Journal of the ACM*, 15(4):514–534.

MUGFORD, L. (1970). A new way of predicting readability. *Reading*, 4:31–35.

MYERS, J., SHINJO, M. et DUFFY, S. (1987). Degree of causal relatedness and memory. *Journal of Memory and Language*, 26(4):453–465.

NEELY, J. (1991). Semantic priming effects in visual word recognition : A selective review of current findings and theories. *In* BESNER, D. et HUMPHREYS, G., éditeurs : *Basic processes in reading : Visual word recognition*, pages 264–. Lawrence Erlbaum Associates Inc., Hillsdale, NJ.

NEW, B., BRYSBAERT, M., VERONIS, J. et PALLIER, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, 28(04):661–677.

NEW, B., PALLIER, C., BRYSBAERT, M. et FERRAND, L. (2004). Lexique 2 : A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, 36(3):516.

NIEVAS, F. et MARI-BEFFA, P. (2002). Negative priming from the non-selected meaning of the homograph. *British Journal of Psychology*, 93(1):47–66.

NUTTALL, C. (1996). *Teaching reading skills in a foreign language*. MacMillan Heinemann, Oxford.

NYMAN, P., KEARL, B. et POWERS, R. (1961). An Attempt to Shorten the Word List with the Dale-Chall Readability Formula. *Educational Research Bulletin*, 40(6):150–152.

O'CONNOR, R., BELL, K., HARTY, K., LARKIN, L., SACKOR, S. et ZIGMOND, N. (2002). Teaching reading to poor readers in the intermediate grades : A comparison of text difficulty. *Journal of Educational Psychology*, 94(3):474–485.

OJEMANN, R. (1934). The reading ability of parents and factors associated with the reading difficulty of parent education materials. *University of Iowa Studies in Child Welfare*, 8:11–32.

OLLER, J. (1972a). Assessing competence in ESL : reading. *TESOL Quarterly*, 6(4):313–323.

OLLER, J. (1972b). Scoring methods and difficulty levels for cloze tests of proficiency in english as a second language. *The Modern Language Journal*, 56(3):151–158.

OLLER, J. et CONRAD, C. (1971). The cloze technique and esl proficiency. *Language Learning*, 21(2):183–194.

OLLER, J. et TULLIUS, J. (1973). Reading Skills of Non-Native Speakers of English. *International Review of Applied Linguistics in Language Teaching*, 11(1):69–80.

O'REGAN, J. et JACOBS, A. (1992). Optimal viewing position effect in word recognition : A challenge to current theory. *Journal of Experimental Psychology : Human Perception and Performance*, 18(1):185–197.

OSTERHOUT, L. et HOLCOMB, P. (1992). Event-related brain potentials elicited by syntactic anomaly* 1. *Journal of memory and language*, 31(6):785–806.

OZASA, T., WEIR, G. et FUKUI, M. (2007). Measuring readability for Japanese learners of English. *In Proceedings of the 12th Conference of Pan-Pacific Association of Applied Linguistics*.

PAGE, M. (2000). Connectionist modelling in psychology : A localist manifesto. *Behavioral and Brain Sciences*, 23(4):443–512.

PAIVIO, A. (1991). Dual coding theory : Retrospect and current status. *Canadian Journal of psychology*, 45(3):255–287.

PAIVIO, A., CLARK, J. et LAMBERT, W. (1988). Bilingual dual-coding theory and semantic repetition effects on recall. *Journal of Experimental Psychology : Learning, Memory, and Cognition*, 14(1):163–172.

PARIZET, M.-L., GRANDET, E. et CORSAIN, M. (2006). *Activités pour le CECR : niveau B1*. CLE International, Paris.

PATTY, W. et PAINTER, W. (1931). A technique for measuring the vocabulary burden of textbooks. *Journal of Educational Research*, 24:127–134.

PAUL, T. (2003). *Guided Independent Reading*. School Renaissance Institute, Madison, WI.

PAUMIER, S. (2003). *De la reconnaissance de formes linguistiques à l'analyse syntaxique*. Thèse de doctorat, Université de Marne-la-Vallée.

PAWLEY, A. et SYDER, F. (1983). Two puzzles for linguistic theory : nativelike selection and nativelike fluency. *In* RICHARDS, J. et SCHMITT, R., éditeurs : *Language and Communication*, pages 191–225. Longman, London.

PEARCE, D. (2002). A comparative evaluation of collocation extraction techniques. *In Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1530–1536.

PERFETTI, C. et HOGABOAM, T. (1975). Relationship between single word decoding and reading comprehension skill. *Journal of Educational Psychology*, 67(4):461.

PERRY, C., ZIEGLER, J. et ZORZI, M. (2007). Nested incremental modeling in the development of computational theories : The CDP+ model of reading aloud. *Psychological review*, 114(2):273–315.

PETERSEN, S., FOX, P., POSNER, M., MINTUN, M. et RAICHLE, M. (1988). Positron emission tomographic studies of the cortical anatomy of single-words processing. *Nature*, 331(6157):585–589.

PETERSEN, S., FOX, P., POSNER, M., MINTUN, M. et RAICHLE, M. (1989). Positron emission tomographic studies of the processing of singe words. *Journal of Cognitive Neuroscience*, 1(2):153–170.

PETERSEN, S. et OSTENDORF, M. (2006a). *A machine learning approach to reading level assessment*. Rapport technique, University of Washington.

PETERSEN, S. et OSTENDORF, M. (2006b). Assessing the reading level of web pages. *In Proceedings of INTERSPEECH 2006 (poster)*, pages 833–836.

PETERSEN, S. et OSTENDORF, M. (2009). A machine learning approach to reading level assessment. *Computer Speech & Language*, 23(1):89–106.

PIENEMANN, M. (1998). *Language processing and second language development : Processability theory*. John Benjamins, Amsterdam.

PIGADA, M. et SCHMITT, N. (2006). Vocabulary acquisition from extensive reading : A case study. *Reading in a Foreign Language*, 18(1):1–28.

PITLER, E. et NENKOVA, A. (2008). Revisiting readability : A unified framework for predicting text quality. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 186–195.

PLATT, J. (1964). Strong Inference. Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science*, 146(3642):347–353.

PLAUT, D. (2005). Connectionist approaches to reading. *In* SNOWLING, M. et HULME, C., éditeurs : *The science of reading : A handbook*, pages 24–38. Blackwell Publishing, Oxford.

POISSON-QUINTON, S. (2005). *Compréhension écrite 2*. CLE International, Paris.

POISSON-QUINTON, S., MAHÉO-LE COADIC, M. et VERGNE-SIRIEYS, A. (2005a). *Festival 1*. CLE International, Paris.

POISSON-QUINTON, S., MAHÉO-LE COADIC, M. et VERGNE-SIRIEYS, A. (2005b). *Festival 3*. CLE International, Paris.

POISSON-QUINTON, S. et MIMRAN, R. (2007). *Compréhension écrite - Niveau 3 B1/B1+*. CLE International, Paris.

POWERS, R., SUMNER, W. et KEARL, B. (1958). A recalculation of four adult readability formulas. *Journal of Educational Psychology*, 49(2):99–105.

PRASAD, R., DINESH, N., LEE, A., MILTSAKAKI, E., ROBALDO, L., JOSHI, A. et WEBBER, B. (2008). The penn discourse treebank 2.0. *In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pages 2961–2968.

PUREN, C. (1988). *Histoire des méthodologies de l'enseignement des langues*. CLE International, Paris.

PUREN, C. (2008). Explication de textes et perspective actionnelle : la littérature entre le dire scolaire et le faire social. *Langage et l'homme*, 43(1):143.

R DEVELOPMENT CORE TEAM (2008). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RABIN, A. (1988). Determining difficulty levels of text written in languages other than English. *In* ZAKALUK, B. et SAMUELS, S., éditeurs : *Readability : Its Past, Present and Future*, pages 46–76. International Reading Association, Newark, Delaware.

RACLE, G. (1988). La lisibilité : quelques aperçus neuropsychologiques. *Communication et langages*, 76(1):20–41.

RAKOTOMALALA, R. (2008). Tests de normalité. Techniques empiriques et tests statistiques. Version 2.0. `http://eric.univ-lyon2.fr/~ricco/cours/cours/Test_Normalite.pdf`.

RAMANAUSKAS, S. (1972). The responsiveness of cloze readability measures to linguistic variables operating over segments of text longer than a sentence. *Reading Research Quarterly*, 8(1):72–91.

RANDALL, M. (2007). *Memory, psychology and second language learning*. John Benjamins Publishing, Amsterdam.

RASCHE, G. (1961). On General Laws and the Meaning of Measurement in Psychology. *In Paper at the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press.

RATNAPARKHI, A. (1996). A maximum entropy model for part-of-speech tagging. *In Proceedings of the conference on empirical methods in natural language processing*, volume 1, pages 133–142.

RAYNER, K. (1986). Eye movements and the perceptual span in beginning and skilled readers. *Journal of Experimental Child Psychology*, 41(2):211–36.

RAYNER, K. (1998). Eye movements in reading and information processing : 20 years of research. *Psychological bulletin*, 124(3):372–422.

RAYNER, K. et DUFFY, S. (1986). Lexical complexity and fixation times in reading : Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, 14(3): 191–201.

RAYNER, K., FOORMAN, B., PERFETTI, C., PESETSKY, D. et SEIDENBERG, M. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest*, 2(2):31–74.

RAYNER, K. et FRAZIER, L. (1989). Selection mechanisms in reading lexically ambiguous words. *Journal of Experimental Psychology : Learning, Memory, and Cognition*, 15(5): 779.

RAYNER, K. et JUHASZ, B. (2006). Reading Processes in Adults. *In* BROWN, K., éditeur : *Encyclopedia of Languages and Linguistics. 2nd Edition*, pages 373–378. Elsevier, Oxford.

RAYNER, K., JUHASZ, B. et POLLATSEK, A. (2005). Eye movements during reading. *In* SNOWLING, M. et HULME, C., éditeurs : *The science of reading : A handbook*, pages 79–99. Blackwell Publishing, Oxford.

RAYNER, K. et POLLATSEK, A. (1989). *The psychology of reading*. Prentice Hall International, Englewood Cliffs, NJ.

REDISH, J. et SELZER, J. (1985). The place of readability formulas in technical communication. *Technical communication*, 32(4):46–52.

ŘEHŮŘEK, R. et SOJKA, P. (2010). Software Framework for Topic Modelling with Large Corpora. *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. `http://is.muni.cz/publication/884893/en`.

REICHER, G. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of experimental psychology*, 81(2):275–280.

REINWEIN, J. (1992). La technique ZigZag comme outil pour mesurer l'effet de l'illustration et du texte sur le lecteur en langue seconde. *In* PREFONTAINE, C. et Lebrun, M., éditeur : *Stratégies d'enseignement et apprentissage en lecture / écriture : actes de colloque*, pages 261–306. Éditions Logiques, Montréal.

REINWEIN, J. (1996). Le temps est-il un indicateur fiable de la difficulté de lecture du texte / de son traitement cognitif par le lecteur ? *Revue québécoise de linguistique*, 25(1):145–162.

REITSMA, P. (1988). Reading practice for beginners : Effects of guided reading, reading-while-listening, and independent reading with computer-based speech feedback. *Reading Research Quarterly*, 23(2):219–235.

RICHARDS, J. et SCHMIDT, R. (2002). Longman dictionary of language teaching and applied linguistics.

RICHAUDEAU, F. (1969). *La lisibilité*. Denoël, Paris.

RICHAUDEAU, F. (1973). *Le langage efficace*. Marabout, Paris.

RICHAUDEAU, F. (1974). 6 phrases, 200 sujets, 42 lapsus, 1 rêve. *Communication et langages*, 23(1):5–24.

RICHAUDEAU, F. (1976). Faut-il brûler les formules de lisibilité ? *Communication et Langages*, 30:6–19.

RICHAUDEAU, F. (1978). Le texte le plus efficace que je connaisse. *Communication et langages*, 37(1):6–24.

RICHAUDEAU, F. (1979). Une nouvelle formule de lisibilité. *Communication et Langages*, 44:5–26.

RICHAUDEAU, F. (1981). Some french work on prose readability and syntax. *Journal of Reading*, 24(6):503–508.

RICHAUDEAU, F. (1984). *Recherches actuelles sur la lisibilité*. Éditions Retz, Paris.

RICHER, J.-J. (2008). Le français sur objectifs spécifiques (f.o.s.) : une didactique spécialisée ? *Synergies Chine*, 3:15–30.

RIGG, P. (1998). The miscue-ESL project. *In* CARRELL, P., DEVINE, J. et ESKEY, D., éditeurs : *Interactive approaches to second language reading*, pages 206–219. Cambridge Univ. Press, Cambridge.

ROGERS, J. (1962). A formula for predicting the comprehension level of material to be presented orally. *The journal of educational research*, 56(4):218–220.

ROSENZWEIG, M. et MCNEILL, D. (1962). Inaccuracies in the semantic count of Lorge and Thorndike. *The American Journal of Psychology*, 75(2):316–319.

ROTHKOPF, E. (1980). Copying span as a measure of the information burden in written language. *Journal of Verbal Learning and Verbal Behavior*, 19(5):562–572.

RUMELHART, D. (1977). Toward an Interactive Model of Reading. *In* DORNIC, S., éditeur : *Attention and Performance VI*, pages 573–603. Lawrence Erlbaum, New York.

RUMELHART, D., HINTON, G. et WILLIAMS, R. (1986a). Learning internal representations by error propagation. *In* RUMELHART, D., MCCLELLAND, J. et the PDP RESEARCH GROUP, éditeurs : *Parallel distributed processing : Explorations in the microstructure of cognition, Vol. 1 : Foundations*, pages 318–364. MIT Press, Cambridge, MA.

RUMELHART, D. et MCCLELLAND, J. (1982). An interactive activation model of context effects in letter perception : II. The contextual enhancement effect and some tests and extensions of the model. *Psychological review*, 89(1):60–94.

RUMELHART, D., MCCLELLAND, J. et the PDP RESEARCH GROUP (1986b). *Parallel distributed processing : Explorations in the microstructure of cognition, Vol. 1 : Foundations*. MIT Press, Cambridge, MA.

SALMELIN, R. et KUJALA, J. (2006). Neural representations of language : activation versus long-range connectivity. *Trends in Cognitive Sciences*, 10(11):519–525.

SAMUELS, S. et KAMIL, M. (1984). Models of the reading process. *In* PEARSON, P., éditeur : *Handbook of reading research*, pages 185–224. Longman, New York.

SANDERS, T. et MAAT, H. (2006). Cohesion and coherence : linguistic approaches. *In* BROWN, K., éditeur : *Encyclopedia of Languages and Linguistics. 2nd Edition*, pages 591–595. Elsevier, Oxford.

SANDERS, T. et NOORDMAN, L. (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse processes*, 29:37–60.

SCARBOROUGH, D., CORTESE, C. et SCARBOROUGH, H. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology : Human Perception and Performance*, 3(1):1–17.

SCARBOROUGH, H. (1990). Index of productive syntax. *Applied Psycholinguistics*, 11(1): 1–22.

SCHLESINGER, I. (1968). *Sentence structure and the reading process*. Mouton, The Hague.

SCHLYTER, S. (2003). Stades de développement en français L2. Exemples d'apprenants suédophones, guidés et non-guidés, du "Corpus Lund".

SCHMID, H. (1994). Probabilistic part-of-speech tagging using decision trees. *In Proceedings of International Conference on New Methods in Language Processing*, volume 12. Manchester, UK.

SCHNEYER, J. (1963). Review : Botel predicting readability levels. *Journal of Developmental Reading*, 7(1):55–57.

SCHULZ, R. (1981). Literature and readability : Bridging the gap in foreign language reading. *Modern Language Journal*, 65(1):43–53.

SCHVANEVELDT, R., ACKERMAN, B. et SEMLEAR, T. (1977). The effect of semantic context on children's word recognition. *Child Development*, pages 612–616.

SCHWANENFLUGEL, P., HARNISHFEGER, K. et STOWE, R. (1988). Context availability and lexical decisions for abstract and concrete words* 1. *Journal of Memory and Language*, 27(5):499–520.

SCHWANENFLUGEL, P. et REY, M. (1986). Interlingual semantic facilitation : Evidence for a common representational system in the bilingual lexicon. *Journal of Memory and Language*, 25(5):605–618.

SCHWARM, S. et OSTENDORF, M. (2005). Reading level assessment using support vector machines and statistical language models. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 523–530.

SCHWARZ, G. (1978). Estimating the dimension of a model. *The annals of statistics*, pages 461–464.

SEARA, R. (2002). L'évolution des méthodologies dans l'enseignement du français langue étrangère depuis la méthodologie traditionnelle jusqu'à nos jours. *Cuadernos del Marqués de San Adrián*, (1):139.

SEBASTIANI, F. (2002). Machine learning in automated text categorization. *ACM computing surveys*, 34(1):1–47.

SEIDENBERG, M. (2006). Connectionist models of reading. *In* GASKELL, G., éditeur : *The Oxford handbook of psycholinguistics*, chapitre 14, pages 235–250. Oxford University Press, New York.

SEIDENBERG, M. et MCCLELLAND, J. (1989). A distributed, developmental model of word recognition and naming. *Psychological review*, 96(4):523–568.

SELVA, T. (2002). Génération automatique d'exercices contextuels de vocabulaire. *In Actes de TALN 2002*, pages 185–194.

SELZER, J. (1981). Readability is a four-letter word. *Journal of business communication*, 18(4):23–34.

SHANAHAN, T., KAMIL, M. et TOBIN, A. (1982). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly*, 17(2):229–255.

SHANNON, C. et WEAVER, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press, Urbana.

SHAPIRO, S. et WILK, M. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611.

SHAW, R. et JACOBSON, M. (1968). A computerized determination of the readability of programmed materials using complete units. *In The annual meeting of the American Educational Research Association*, Chicago.

SI, L. et CALLAN, J. (2001). A statistical model for scientific readability. *In Proceedings of the Tenth International Conference on Information and Knowledge Management*, pages 574–576. ACM New York, NY, USA.

SILVA, J., DIAS, G., GUILLORÉ, S. et PEREIRA LOPES, J. (1999). Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. *Progress in Artificial Intelligence*, pages 849–849.

SILVA, J. et LOPES, G. (1999). A local maxima method and a fair dispersion normalization for extracting multi-word units from corpora. *In Sixth Meeting on Mathematics of Language*.

SILVERBERG, S. et SAMUEL, A. (2004). The effect of age of second language acquisition on the representation and processing of second language words. *Journal of memory and language*, 51(3):381–398.

SINGER, H. (1975). The seer technique : A non-computational procedure for quickly estimating readability level. *Journal of Literacy Research*, 7(3):255–267.

SLATER, W., GRAVES, M., SCOTT, S. et REDD-BOYD, T. (1988). Discourse structure and college freshmen's recall and production of expository text. *Research in the Teaching of English*, 22(1):45–61.

SMADJA, F. (1993). Retrieving collocations from text : Xtract. *Computational linguistics*, 19(1):143–177.

SMILEY, S., OAKLEY, D., WORTHEN, D., CAMPIONE, J. et BROWN, A. (1977). Recall of thematically relevant material by adolescent good and poor readers as a function of written versus oral presentation. *Journal of Educational Psychology*, 69(4):381–387.

SMITH, E. (1961). Devereaux readability index. *The Journal of Educational Research*, 54(8):289–303.

SMITH, E. et SENTER, R. (1967). Automated Readability Index. Rapport technique, AMRL-TR-66-220, Aerospace Medical Research Laboratories, Wright-Patterson Airforce Base, OH.

SMITH, F. (1969). Familiarity of configuration vs. discriminability of features in the visual identification of words. *Psychonomic Science*, 14:261–262.

SMITH, F. (1971). *Understanding reading*. Holt, Rinehart and Winston, New York.

SNOW, C. et HOEFNAGEL-HÖHLE, M. (1978). The critical period for language acquisition : Evidence from second language learning. *Child development*, 49(4):1114–1128.

SNOWDON, D., KEMPER, S., MORTIMER, J., GREINER, L., WEKSTEIN, D. et MARKESBERY, W. (1996). Linguistic ability in early life and cognitive function and Alzheimer's disease in late life. *Journal of the American Medical Association*, 275(7):528–532.

SOLOMON, R. et POSTMAN, L. (1952). Frequency of usage as a determinant of recognition thresholds for words. *Journal of Experimental Psychology*, 43(3):195–201.

SPACHE, G. (1953). A new readability formula for primary-grade reading materials. *The Elementary School Journal*, 53(7):410–413.

SPACHE, G. (1960). *Good Reading For Poor Readers*. The Garrard Press, Champaign : Illinois.

SPARCK JONES, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

SPAULDING, S. (1951). Two formulas for estimating the reading difficulty of Spanish. *Educational Research Bulletin*, 30(5):117–124.

SPAULDING, S. (1956). A Spanish readability formula. *Modern Language Journal*, 40(8): 433–441.

SPEARMAN, C. (1904). "General Intelligence", objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292.

SPERBERG-MCQUEEN, C., BURNARD, L. *et al.* (1994). *Guidelines for Electronic Text Encoding and Interchange*. Citeseer, Chicago and Oxford.

SPIVEY, N. (1987). Construing constructivism : : Reading research in the United States. *Poetics*, 16(2):169–192.

STADTHAGEN-GONZALEZ, H. et DAVIS, C. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, 38(4):598–605.

STANOVICH, K. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading research quarterly*, 16(1):32–71.

STEFFENSEN, M., JOAG-DEV, C. et ANDERSON, R. (1979). A cross-cultural perspective on reading comprehension. *Reading research quarterly*, pages 10–29.

STEINEL, M., HULSTIJN, J. et STEINEL, W. (2007). Second language idiom learning in a paired-associate paradigm : Effects of direction of learning, direction of testing, idiom imageability, and idiom transparency. *Studies in Second Language Acquisition*, 29(3): 449–484.

STENNER, A. (1996). Measuring reading comprehension with the lexile framework. *In Fourth North American Conference on Adolescent/Adult Literacy*.

STENNER, A. et BURDICK, D. (1997). *The Objective Measurement of Reading Comprehension : In Response to Technical Questions Raised by the California Department of Education Technical Study Group*. Rapport technique, MetaMetrics, Inc., Durham, NC.

STENNER, A., HORADIN, I., SMITH, D. et SMITH, R. (1988). *The Lexile Framework*. MetaMetrics, Inc., Durham, NC.

STENNER, A., SMITH, M. et BURDICK, D. (1983). Toward a theory of construct definition. *Journal of Educational Measurement*, 20(4):305–316.

STEVENS, K. (1980). Readability formulae and McCall-Crabbs standard test lessons in reading. *The Reading Teacher*, 33(4):413–415.

STEVENS, S. (1946). On the theory of scales of measurement. *Science*, 103(2684):677–680.

STEVENS, S. (1968). Measurement, statistics, and the schemapiric view. *Science*, 161(3844):849–856.

STEVENSON, D. (1981). 'all of the above' : On problems in the testing of fl reading. *System*, 9(3):267–273.

STOCKER, L. (1971). Increasing the Precision of the Dale-Chall Readability Formula. *Reading Improvement*, 8(3):87–89.

STONE, C. (1956). Measuring difficulty of primary reading material : A constructive criticism of Spache's measure. *The Elementary School Journal*, 57(1):36–41.

STONE, M. et MULLINS, M. (1932). A Study of the Difficulty of Vocabulary and Sentence Structure of Sixteen French Selections. *French Review*, 6(1):15–36.

SZALAY, T. (1965). Validation of the Coleman readability formulas. *Psychological Reports*, 17(3):965–966.

TABOSSI, P. et ZARDON, F. (1993). Processing ambiguous words in context. *Journal of Memory and Language*, 32(3):359–372.

TAN, P.-N., STEINBACH, M. et KUMAR, V. (2005). *Introduction to Data Mining*. Addison-Wesley, Boston.

TANAKA-ISHII, K., TEZUKA, S. et TERADA, H. (2010). Sorting texts by readability. *Computational Linguistics*, 36(2):203–227.

TARDIF, J. (1990). Les retombées de la psychologie cognitive dans les recherches en lecture. *In* BOYER, J.-Y. et LEBRUN, M., éditeurs : *L'actualité de la recherche en lecture*, pages 9–31. ACFAS, Montréal.

TAYLOR, B. (1980). Children's memory for expository text after reading. *Reading Research Quarterly*, 15(3):399–411.

TAYLOR, W. (1953). Cloze procedure : A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

TAYLOR, W. (1957). "Cloze" readability scores as indices of individual differences in comprehension and aptitude. *Journal of Applied Psychology*, 41(1):19–26.

TENPENNY, P. (1995). Abstractionist versus episodic theories of repetition priming and word identification. *Psychonomic Bulletin & Review*, 2(3):339–363.

THARP, J. (1939). The Measurement of Vocabulary Difficulty. *Modern Language Journal*, pages 169–178.

THERNEAU, T. et ATKINSON, E. (1997). An introduction to recursive partitioning using the rpart routines. Rapport technique, Mayo Foundation.

THORNDIKE, E. (1921). Word knowledge in the elementary school. *The Teachers College Record*, 22(4):334–370.

TIMBAL-DUCLAUX, L. (1984). La transparence du texte pour mesurer sa lisibilité. *Communication et langages*, 59(1):9–20.

TIMBAL-DUCLAUX, L. (1985). Textes "inlisable" et lisible. *Communication et Langages*, 66:13–31.

TIMBAL-DUCLAUX, L. (1989). Les 4 couleurs du texte. *Communication et langages*, 80(1):5–26.

TRACEY, D. et MORROW, L. (2006). *Lenses on reading : An introduction to theories and models*. The Guilford Press, New York.

TREMBLAY, G. (1984). Nos journaux sont-ils compréhensibles. *Revue canadienne de science politique*, 17(3):595–605.

TROLLINGER, W. et KAESTLE, C. (1986). *Difficulty of Text as a Factor in the History of Reading*. Rapport technique, Program Report 86-13, Wisconsin Center for Education Research, Madison.

TSCHICHOLD, C. (2006). Intelligent call : The magnitude of the task. *In Actes de la 13e Conféerence sur le Traitement Automatique des Langues Naturelles (TALN2006)*.

TUFFÉRY, S. (2007). *Data mining et statistique décisionnelle l'intelligence des données*. Éd. Technip, Paris.

UITDENBOGERD, A. (2003). Using the web as a source of graded reading material for language acquisition. *International Conference on Web-based Learning*, pages 423–432.

UITDENBOGERD, A. (2010). Fun with Filtering French. *In Proceedings of Australasian Language Technology Association Workshop*, pages 58–66.

UITDENBOGERD, S. (2005). Readability of French as a foreign language and its uses. *In Proceedings of the Australian Document Computing Symposium*, pages 19–25.

UNDERWOOD, G., SCHMITT, N. et GALPIN, A. (2004). The eyes have it : An eye-movement study into the processing of formulaic sequences. *In* SCHMITT, N., éditeur : *Formulaic sequences : acquisition processing and use*, pages 155–172. John Benjamins, Amsterdam.

UNDERWOOD, N. et MCCONKIE, G. (1985). Perceptual span for letter distinctions during reading. *Reading Research Quarterly*, 20:153–162.

URQUHART, A. et WEIR, C. (1998). *Reading in a second language : Process, product and practice*. Longman, London.

VAN DIJK, T. (1980). *Text and context : Explorations in the semantics and pragmatics of discourse*. Longman, London.

van HELL, J. et de GROOT, M. (1998). Disentangling Context Availability and Concreteness in Lexical Decision and Word Translation. *The Quarterly Journal of Experimental Psychology, Section A : Human Experimental Psychology*, 51(1):41–63.

VAN HELL, J. et MAHN, A. (1997). Keyword mnemonics versus rote rehearsal : Learning concrete and abstract foreign words by experienced and inexperienced learners. *Language Learning*, 47(3):507–546.

van HELL, J. et TOKOWICZ, N. (2010). Event-related brain potentials and second language learning : Syntactic processing in late L2 learners at different L2 proficiency levels. *Second Language Research*, 26(1):43–74.

VAN HEUVEN, W. (2005). Bilingual interactive activation models of word recognition in a second language. *In* COOK, V. et BASSETTI, B., éditeurs : *Second Language Writing Systems*, pages 260–288. Multilingual Matters, Clevedon.

VAN HEUVEN, W., DIJKSTRA, T. et GRAINGER, J. (1998). Orthographic neighborhood effects in bilingual word recognition. *Journal of Memory and Language*, 39(3):458–483.

van OOSTEN, P. et HOSTE, V. (2011). Readability Annotation : Replacing the Expert by the Crowd. *In Sixth Workshop on Innovative Use of NLP for Building Educational Applications*.

van OOSTEN, P., HOSTE, V. et TANGHE, D. (2011). A posteriori agreement as a quality measure for readability prediction systems. *In* GELBUKH, A., éditeur : *Computational Linguistics and Intelligent Text Processing*, volume 6609 de *Lecture Notes in Computer Science*, pages 424–435. Springer, Berlin / Heidelberg.

VAN ORDEN, G. (1987). A rows is a rose : Spelling, sound, and reading. *Memory & Cognition*, 15:181–198.

VAN ORDEN, G., JOHNSTON, J. et HALE, B. (1988). Word identification in reading proceeds from spelling to sound to meaning. *Journal of Experimental Psychology : Learning, Memory, and Cognition*, 14:371–386.

VAN ORDEN, G., PENNINGTON, B. et STONE, G. (1990). Word identification in reading and the promise of subsymbolic psycholinguistics. *Psychological Review*, 97(4):488–522.

VAN ROEY, J. (1990). *French-English contrastive lexicology : An Introduction*. Peeters Publishers, Louvain-la-Neuve.

VELLEMAN, P. et WILKINSON, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *American Statistician*, 47(1):65–72.

VENABLES, W. et RIPLEY, B. (2002). *Modern Applied Statistics with S*. Springer, New York.

VENEZKY, R. (1984). The history of reading research. *In* PEARSON, P., éditeur : *Handbook of reading research*, pages 3–38. Longman, New York.

VERLINDE, S., SELVA, T. et BINON, J. (2003). Alfalex : un environnement d'apprentisage du vocabulaire français en ligne, interactif et automatisé. *Romaneske*, 28(1):42–62.

VIGNER, G. (1979). *Lire, du texte au sens*. CLE International, Paris.

VITU, F., O'REGAN, J. et MITTAU, M. (1990). Optimal landing position in reading isolated words and continuous text. *Perception & Psychophysics*, 47(6):583–600.

VLAD, M. (2006). *Lire des textes en français langue étrangère à l'école*. E.M.E. / Inter-Communications, Cortil-Wodon.

VOGEL, M. et WASHBURNE, C. (1928). An objective method of determining grade placement of children's reading material. *The Elementary School Journal*, 28(5):373–381.

WAN, X., LI, H. et XIAO, J. (2010). EUSUM : extracting easy-to-understand english summaries for non-native readers. *In Proceeding of the 33rd international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 491–498.

WASHBURNE, C. et MORPHETT, M. (1938). Grade placement of children's books. *The Elementary School Journal*, 38(5):355–364.

WATRIN, P. et FRANÇOIS, T. (2011). N-gram frequency database reference to handle MWE extraction in NLP applications. *In Proceedings of the 2011 Workshop on MultiWord Expressions : from Parsing and Generation to the Real World (ACL Workshop)*, pages 83–91.

WEAVER, B. (2000). *Leveling books K–8 : Matching readers to text*. International Reading Association, Newark, DE.

WEAVER, W. et KINGSTON, A. (1963). A factor analysis of the cloze procedure and other measures of reading and language ability. *Journal of Communication*, 13(4):252–261.

WEBER, R.-M. (1970). First-Graders'Use of Grammatical Context in Reading. *In* LEVIN, H. et WILLIAMS, J., éditeurs : *Basic Studies on Reading*, pages 147–163. Basic Book, New York.

WEBER-FOX, C. et NEVILLE, H. (1996). Maturational constraints on functional specializations for language processing : ERP and behavioral evidence in bilingual speakers. *Journal of cognitive neuroscience*, 8(3):231–256.

WEIR, G. et ANAGNOSTOU, N. (2008). Collocation frequency as a readability factor. *In Proceedings of the 13th Conference of the Pan Pacific Association of Applied Linguistics*.

WEIR, G. et RITCHIE, C. (2006). Estimating readability with the Strathclyde readability measure. *In ICTATLL Workshop Preprints*, pages 21–22.

WELSCH, R. (2006). 20 Ways To... Increase Oral Reading Fluency. *Intervention in School and Clinic*, 41(3):180–183.

WEST, R. et STANOVICH, K. (1978). Automatic contextual facilitation in readers of three ages. *Child Development*, pages 717–727.

WHALEY, C. (1978). Word–nonword classification time. *Journal of Verbal Learning and Verbal Behavior*, 17(2):143–154.

WHEELER, L. et SMITH, E. (1954). A practical readability formula for the classroom teacher in the primary grades. *Elementary English*, 31:397–99.

WHEELER, L. et WHEELER, V. (1948). Selecting appropriate reading materials. *Elementary English*, 25:478–89.

WILLMS, J. (2003). Literacy proficiency of youth : Evidence of converging socioeconomic gradients. *International Journal of Educational Research*, 39(3):247–252.

WINKIN, Y. (1981). *La nouvelle communication*. Editions du Seuil, Paris.

WITTEN, I. et FRANK, E. (2005). *Data Mining : Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco.

ZAKALUK, B. et SAMUELS, S. (1988). Toward a New Approach to Predicting Text Comprehensibility. *In* ZAKALUK, B. et SAMUELS, S., éditeurs : *Readability : Its Past, Present and Future*, pages 121–144. International Reading Association, Newark, Delaware.

ZAKALUK, B. et SAMUELS, S. (1996). Issues related to text comprehensibility : the future of readability. *Revue québécoise de linguistique*, 25(1):41–59.

ZEVIN, J. et SEIDENBERG, M. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language*, 47(1):1–29.

ZIPF, G. (1965). *The psycho-biology of language*. MIT Press, Cambridge, MA.