

Le traitement automatique du langage pour l'évaluation automatique de la difficulté lexicale en FLE

FRANÇOIS, Thomas

(Université catholique de Louvain, IL&C, Cental)

Résumé

Récemment, le développement des nouvelles technologies, en particulier celles dédiées au traitement automatique du langage, a ouvert de nouvelles perspectives pour la didactique du français langue étrangère. Dans cet article, nous proposons d'en présenter plusieurs, qui ont toutes en commun le même objectif : l'évaluation automatisée de la difficulté lexicale. Tout d'abord, les formules de lisibilité sont capables d'analyser automatiquement un texte et d'en prédire globalement la difficulté à la lecture. Nous discutons ce qu'elles peuvent apporter au FLE et présentons une nouvelle formule spécialisée pour le FLE et reposant sur la technologie de l'apprentissage profond. Dans un second temps, nous nous focalisons plus particulièrement sur le lexique au travers du projet CEFRLex, qui propose, dans différentes langues, des ressources lexicales incluant la distribution de fréquence de mots sur les niveaux du cadre européen commun de référence pour les langues (CECR). Enfin, nous discutons comment ces informations fréquentielles peuvent être mises à profit en vue de prédire le niveau CECR des mots dans un texte.

近年、新しい科学技術、とりわけ自動言語処理技術の発展により外国語としてのフランス語教育の新たな可能性が広がっている。本論文は、その技術を活用した語彙難度の自動評価に関する取り組みをいくつか紹介するものである。まず、文章を自動で分析し全体の可読性を予測することができるリーダビリティ公式について記述する。これによる FLE への貢献について議論し、ディープラーニング技術に基づく FLE に特化した新たな公式を提示したい。次に、複数言語を対象に、ヨーロッパ言語共通参照枠 (CECR) のレベルに応じた語彙頻度情報等の語彙リソースを提供する CEFRLex プロジェクトを通して、さらに語彙に焦点を当てた議論を行う。最後に、この頻度情報を用いて、テキスト中の単語の CEFR レベルを予測する方法について述べる。

Mots clés

Traitement automatique du langage, FLE, Lisibilité, Difficulté lexicale, CEFRLex

自動言語処理, FLE, リーダビリティ, 語彙難度, CEFRLex

1 Introduction

L'évaluation de la difficulté du vocabulaire, notamment dans le contexte du français langue étrangère (FLE), est une préoccupation relativement ancienne. Elle est déjà attestée dans un article de Tharp (1939), qui rapporte non seulement l'existence de listes de fréquence (par exemple, le *French Word Book* de Henmomm, qui contenait 3900 mots dont les fréquences avaient été estimées sur 400 000 mots, ou la liste de Vander Beke, 1932), mais aussi la volonté d'évaluer la lisibilité de livres simplifiées en français au moyen du lexique. Ces outils visaient à standardiser la sélection des contenus mis à disposition des apprenants afin de la rendre plus efficace. Néanmoins, ils péchaient par leur lourdeur : il était nécessaire de vérifier manuellement la présence de chaque mot du texte dans la liste de référence afin d'en connaître la fréquence.

Depuis lors, de nombreuses évolutions ont eu lieu dans le domaine de la mesure de la difficulté des textes en langue étrangère (L2). Tout d'abord, l'émergence, puis le développement fulgurant des ordinateurs a permis d'automatiser diverses opérations qui étaient auparavant extrêmement consommatrices de temps : par exemple, le comptage des mots dans un corpus, l'estimation de fréquences, la consultation d'une liste de mots simples, ou encore l'extraction de variables capturant les différentes dimensions constitutives de la difficulté d'un mot telles que sa longueur, sa structure vocalique, ses constituants morphologiques, etc. (François *et al.*, 2016). Cette évolution a été rendue plus aisée encore grâce aux technologies issues du traitement automatique du langage, un domaine à la croisée de la linguistique et de l'intelligence artificielle, qui vise à traiter, analyser ou générer automatiquement des données langagières¹. Appliquées au domaine de l'enseignement des langues étrangères, ces technologies ont conduit au développement de l'ALiAO ou l'apprentissage des langues intelligemment assistés par ordinateur, lequel permet aujourd'hui d'envisager d'automatiser diverses tâches pédagogiques telles que la correction automatique des erreurs des apprenants, la production de feedback, la recherche de documents authentiques adaptés au niveau de l'apprenant, la pratique de la lecture, ou encore la génération automatique d'exercices.

La problématique de l'évaluation de la difficulté lexicale a également connu de notables évolutions sous l'effet des nouvelles technologies depuis l'époque de Tharp (1939). Cette tâche a,

¹ Pour une introduction à ce domaine, nous renvoyons à l'ouvrage de synthèse de Jurafsky et Martin (2009).

pendant longtemps, été liée au domaine de la lisibilité, qui vise à mesurer la facilité de lecture de textes à l'aide de méthodes reproductibles. Aujourd'hui, on trouve également des travaux qui se focalisent sur la détection automatique des mots complexes (Tack, 2021).

Dans cet article, qui s'inspire d'une communication donnée en 2021 à Kyoto et portant le même titre, nous nous proposons de présenter différentes pistes concrètes pour l'évaluation automatique de la difficulté lexicale en FLE que nous avons explorées ces dernières années. La section 2 traite de l'approche holistique de la difficulté lexicale à l'aide de formules de lisibilité. Après avoir brièvement décrit les travaux les plus notables à ce niveau, nous présenterons une nouvelle formule de lisibilité, spécifique au FLE et reposant sur des technologies d'apprentissage profond (*deep learning*). Les sections 3 et 4 aborderont l'usage de ressources graduées, dans le cadre du projet CEFRLex, pour évaluer la difficulté au niveau du mot. La section 3 décrira les principes théoriques et méthodologiques ayant présidé à la constitution de ce type de ressources, tandis que la section 4 présentera diverses façons d'exploiter ce type de ressources afin de prédire automatiquement la difficulté des mots pour les apprenants de FLE.

2 La lisibilité des textes de FLE

2.1 Les travaux en lisibilité

Les origines des premières formules de lisibilité remontent au début du 20^{ème} siècle aux États-Unis et, en particulier, à l'étude séminale de Lively et Pressey (1923). Il est intéressant de noter que cette première approche de la mesure de la difficulté de lecture des textes se concentre sur les effets du lexique. C'est ensuite avec Vogel et Washburne (1928) que l'effet de la syntaxe sur la lecture est également pris en compte. Par la suite, de nombreuses formules de lisibilité sont développées. Elles suivent toutes la même méthodologie : sur la base d'un ensemble de textes dont le niveau de difficulté à la lecture a été calibré, on entraîne une régression linéaire multiple dont les variables capturent les propriétés des textes à analyser. Les formules les plus connues, encore aujourd'hui, sont celles de la période classique, à savoir celles de Flesch (1948), de Dale et Chall (1948) et de Gunning (1952). Par la suite, les formules commencent à s'adapter à l'automatisation (Smith et Senter, 1967). Dans le dernier quart du siècle, le développement de la grammaire du texte et de la psychologie cognitive enrichit les formules, via la prise en compte de phénomènes textuels tels que la cohésion, l'inférence, etc.²

À terme, ce qui aura le plus d'impact sur la conception des formules lisibilité est, comme

² Pour un panorama détaillé des principales formules de lisibilité du 20^e siècle, consulter la synthèse de François (2011).

rappelé en introduction, le développement de l'informatique et de l'intelligence artificielle. À l'aube du 21^e siècle, divers travaux rompent avec l'usage de la régression linéaire à laquelle ils préfèrent des algorithmes d'apprentissage automatisé, comme un classifieur Bayésien naïf (Collins-Thompson et Callan, 2004), des machines à vecteur support (Schwarm et Ostendorf, 2005), la régression logistique (Heilman *et al.*, 2008), les arbres de décision (Kate *et al.*, 2010) ou le *boosting* (François et Fairon, 2012). Au niveau des propriétés textuelles capturées dans des variables, les technologies du TAL permettent de prendre en compte bien davantage de phénomènes linguistiques et d'une manière moins superficielle³. François et Miltakaki (2012) montreront ainsi que la combinaison de variables classiques en lisibilité – comme la longueur des phrases, des mots ou la fréquence lexicale – et de variables basées sur le TAL – comme la détection des flexions verbales, les propriétés des expressions polylexicales ou la fréquence de bigrammes – permettent de dépasser les limites des formules classiques.

Récemment, le domaine de la lisibilité n'a pas échappé à la révolution neuronale et les modèles de lisibilité les plus récents reposent sur des réseaux de neurones profonds (Martinc *et al.*, 2021 ; Nadeem et Ostendorf, 2018) ainsi que sur des représentations riches de la langue via les plongements de mots (Cha *et al.*, 2017 ; Filighera *et al.*, 2019)⁴. Les performances atteintes par ces modèles aujourd'hui laissent envisager un usage plus massif de ce type d'outils dans le cadre de l'ALiAO.

2.1 Une nouvelle formule neuronale pour le FLE

La grande majorité des formules de lisibilité ont été conçues pour des enfants ou des adultes éprouvant des difficultés de lecture, mais il est difficile de déterminer quand la première formule spécialisée pour les apprenants d'une L2 est apparue. Cela s'explique par le fait que, jusque dans les années 70, le processus de lecture était considéré comme universel et peu sensible aux variations de population. Nous savons toutefois aujourd'hui que les apprenants d'une L2 se différencient des natifs à plusieurs égards (Koda, 2005 : 7) : (1) ils doivent gérer des interférences de leur L1 ; (2) ils sont généralement plus âgés que les enfants natifs et disposent d'un meilleur réseau conceptuel et (3) ils ne peuvent pas s'appuyer sur une connaissance linguistique acquise à l'oral comme les enfants natifs. Il semble donc logique de développer des formules de lisibilité spécifique aux apprenants d'une L2.

Pourtant, pendant longtemps, il était d'usage de recourir à une formule L1, par exemple celle de Flesch (1948), pour analyser des textes à destination d'apprenants de l'anglais L2. De la même

³ Cette période est résumée dans les synthèses suivantes : Collins-Thompson (2014), François (2015).

⁴ Consulter Vajjala (2022) pour une synthèse récente qui inclut les travaux neuronaux.

manière, Cornaire (1988) considéra que la formule de Henry (1975), développée pour des étudiants de primaire et de secondaire, était adéquate pour des apprenants de FLE. Cette question n'a pas encore été totalement tranchée expérimentalement, mais on considère généralement aujourd'hui qu'il est préférable de spécialiser les formules de lisibilité, y compris pour les lecteurs L2. Pour le FLE, cet idéal reste difficile à mettre en œuvre, par manque de ressources. On ne compte guère en effet que les modèles de Uitdenbogerd (2005) et de François et Fairon (2012). Le second représentait le modèle « état de l'art » jusque récemment, présentant la capacité de classer correctement 49 % des textes selon l'échelle du cadre européen commun de référence pour les langues (CECR). Néanmoins, aucun de ces deux modèles n'est disponible pour la recherche ou l'enseignement en FLE.

C'est pour cette raison que, récemment, nous avons mis au point une nouvelle formule de lisibilité pour le FLE (Yancey *et al.*, 2021), qui propose plusieurs avancées sur l'existant. Tout d'abord, elle repose sur les dernières technologies dans le domaine de l'intelligence artificielle, à savoir le *Deep Learning* ou apprentissage profond. Cette technologie utilise les réseaux de neurones pour apprendre automatiquement la relation complexe qui existe entre le niveau de difficulté d'un texte et une représentation riche de ce texte à l'aide de plongements de mots de BERT (Devlin *et al.*, 2018). BERT est un exemplaire de ce qu'on appelle des modèles de langue : construit à partir d'un énorme corpus de textes, ce type de modèle permet de représenter la langue (et par extension, des textes) à l'aide d'un nombre important de dimensions codant des propriétés sémantiques, mais aussi syntaxiques, morphologiques, etc.

Le second atout de ce modèle de lisibilité pour le FLE est la taille de son jeu d'entraînement, puisque pas moins de 2751 textes, issus de 48 manuels de FLE différents et annotés suivant l'échelle du CECR, ont été utilisés pour affiner le modèle BERT (*fine-tuning*) en le spécialisant à la tâche d'évaluation de la lisibilité pour le FLE.

Une troisième contribution de ce travail est le fait que la formule sera disponible sur le site DMesure⁵, qui est gratuitement accessible à tous les chercheurs et enseignants de FLE intéressés d'évaluer des textes selon l'échelle du CECR. Les performances de ce nouveau modèle atteignent les 58 % de textes correctement classés, avec seulement 10 % d'erreurs graves (définie comme une prédiction s'écartant de plus d'un niveau CECR par rapport à la valeur de référence), ce qui le rend utilisable dans des contextes réels, même s'il convient de rester prudent concernant les résultats. Dans cette optique, l'interface, dont la version actuelle est présentée à la Figure 1, propose deux types d'information propres à aider l'utilisateur à mieux appréhender les prédictions réalisées par

⁵ À l'adresse suivante : <https://cental.uclouvain.be/dmesure/>

le système. D'une part, comme BERT produit les probabilités que le texte analysé appartienne à chacun des 6 niveaux du CECR (les niveaux C1 et C2 étant conjoints), le système les affiche afin que l'utilisateur puisse voir si le modèle hésite ou non sur le niveau à prédire. D'autre part, le système calcule également – en se basant sur FABRA (Wilkens *et al.*, 2022) – un ensemble de variables linguistiques connues pour être corrélées à la difficulté de lecture (ex. la longueur des mots, leur familiarité, la proportion de propositions relatives, etc.) et les affiche à l'écran. Ces informations permettent à l'utilisateur de mieux comprendre quelles sont les caractéristiques qui participent le plus au diagnostic général du modèle.

Niveau le plus probable pour ce texte : B1

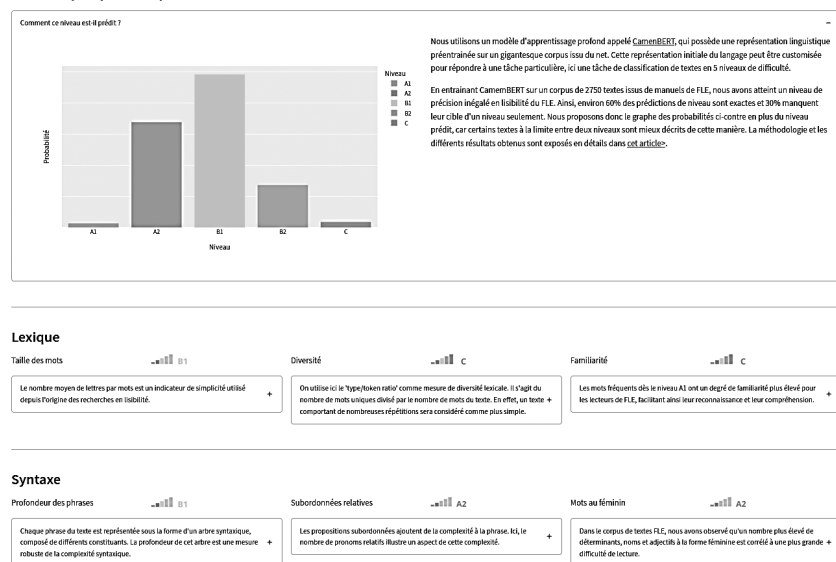


Figure 1: capture d'écran du site DMeasures (en construction)

3 Le projet CEFRLex

Le projet CEFRLex est constitué d'un ensemble de ressources dans plusieurs langues européennes qui présentent, pour chaque mot, une distribution de fréquence sur les 6 niveaux du CECR. Ces ressources se distinguent en cela des listes de fréquence classiques rapportant une fréquence globale, estimée sur un large corpus. Dans le projet CEFRLex, les fréquences sont estimées sur un corpus de manuels pédagogiques et de livres simplifiés assignés à un niveau donné du CECR. Les données sont accessibles gratuitement en ligne et téléchargeable pour la recherche

et l'enseignement⁶.

3.1 Motivations pour le projet CEFRLex

Comme nous l'avons discuté précédemment, l'une des principales ressources à la base de l'évaluation de la difficulté lexicale est la liste de fréquence. Depuis Thorndike (1921), qui passa dix ans à compter les fréquences de mots anglais dans un corpus de 4,5 millions de mots afin d'établir la première liste de fréquence d'importance pour l'anglais, il est devenu habituel de se reposer sur de telles listes pour évaluer la difficulté des mots. Cette pratique se justifie par la corrélation élevée existant entre la fréquence des mots et leur difficulté (Howes et Salomon, 1951), ce que l'on désigne comme l'effet de fréquence. Néanmoins, cet effet peut varier sensiblement en fonction du type de textes utilisés pour créer la liste de fréquence. La plupart des listes développées pour l'anglais ou le français – par exemple, pour le français, les listes de Gougenheim *et al.* (1964) ; New *et al.* (2004) ou Lonsdale et Le Bras (2009) – incluent des textes à destination de locuteurs natifs, qui restent bien souvent encore l'horizon à atteindre pour les apprenants. Or, Brysbaert *et al.* (2021 : 226) ont comparé le degré de connaissance, par des apprenants L2 de l'anglais, de 500 mots – collectés par Hashimoto et Egbert (2019) – avec les rangs basés sur des fréquences L1 et L2. Ils en concluent que les connaissances lexicales en L1 et en L2 sont sensiblement différentes. C'est pourquoi, dans le projet CEFRLex, les corpus servant à estimer les fréquences sont composés de textes destinés à des apprenants L2 et donc plus proches de la connaissance lexicale réelle des apprenants.

Une seconde limite de l'utilisation de listes de fréquence classiques découle des remarques précédentes, à savoir qu'elles postulent qu'un mot se caractérise par une fréquence uniforme tout au long de l'apprentissage. Ainsi, le mot « absolument » est classé à la 235^e position dans Gougenheim, avec une fréquence de 143 occurrences. Il apparaît donc comme un mot fréquent à tout moment de l'apprentissage du FLE, qui gagnerait dès lors à être enseigné dès le niveau A1. Pourtant, lorsqu'on le dénombre dans les manuels de FLE, on peut observer qu'il est quasiment absent des manuels A1 et apparaît ensuite clairement à partir de A2. Il s'agit donc d'un mot fréquent, mais à partir d'un certain moment de l'apprentissage. C'est pour mieux dresser ces profils lexicaux que les ressources de CEFRLex présentent la distribution de fréquence des mots sur les 6 niveaux du CECR⁷.

Les ressources du projet CEFRLex se positionnent également par rapport à un autre type de ressources pédagogiques : les référentiels pour le CECR. Développés pour plusieurs langues,

⁶ L'adresse du projet est : <https://cental.uclouvain.be/cefrlex/>

⁷ Pour certaines langues, le niveau C2 n'a pas pu être modélisé, par manque de données de ce niveau.

ceux-ci proposent un inventaire de formes linguistiques (mots, structures syntaxiques, graphèmes, etc.) ou de fonctions à acquérir à chaque niveau du CECR. Pour le français, les référentiels ont été développés par Jean-Claude Beacco en collaboration avec France Éducation Internationale, de 2004 à 2016 et seulement pour les niveaux A1 à B2. En ce qui concerne le lexique, il est listé aux chapitres 4 (notions générales) et 6 (notions spécifiques) et comprend environ 8 500 entrées, dont près de 3000 sont polysémiques. L'assignation d'un mot à un niveau CECR a été effectué sur des critères divers : les descriptions officielles des descripteurs du CECR, l'expertise des chercheurs et enseignants participant au projet et des exemples de productions d'apprenants.

Les référentiels pour le CECR présentent toutefois également certaines limitations. D'une part, Hulstijn (2007) a mis en question les critères utilisés pour établir le niveau CECR de chaque mot, mettant en avant le manque d'empirisme dans le processus de création. D'autre part, ce type de ressources ne fait aucune différence entre les mots au sein d'un même niveau. Or, des niveaux comme le B1 ou le B2 correspondent à de nombreuses heures d'apprentissage durant lesquels il est nécessaire de phaser l'apprentissage du lexique. À ce niveau, les distributions du projet CEFRLex offre une meilleure granularité pour réaliser cet objectif pédagogique.

3.3 FLELex : méthodologie de conception

Actuellement, il existe 7 ressources différentes au sein du projet CEFRLex. Celle qui nous intéresse ici, puisqu'elle concerne le français, est FLELex (François *et al.*, 2014). Par la suite, la même méthodologie a été adaptée à 5 autres langues (suédois, anglais, néerlandais, allemand et espagnol). Cette méthodologie est la suivante : (1) il s'agit de collecter un corpus de textes pédagogiques annotés, aussi fidèlement que possible, selon l'échelle du CECR ; c'est pourquoi nous avons utilisé soit des manuels d'apprentissage, soit des livres simplifiés. (2) Le corpus doit ensuite être analysé automatiquement à l'aide de programmes informatiques capables d'identifier le lemme et la catégorie de discours de chaque mot du corpus. (3) Une fois le corpus prêt, le décompte de chaque lemme est réalisé automatiquement et les fréquences obtenus sont normalisées. Enfin, (4) la ressource ainsi obtenue est nettoyée à l'aide d'un processus itératif semi-manuel. Nous détaillons ci-après chacune de ces quatre étapes.

En ce qui concerne le corpus, celui-ci comportait 28 manuels de FLE publiés après 2001 dans lesquels seuls des textes associés à des tâches de compréhension à la lecture ont été retenus. Par conséquent, la ressource est adéquate pour évaluer la fréquence des mots du FLE dans un contexte de réception. En complément de ces manuels, 29 livres simplifiés ont été intégrés au corpus, pour un total de 2071 textes et de 777 000 mots. Divers genres de textes sont également représentés : textes narratifs, informatifs, dialogues, lettres, e-mails, chansons ou encore des recettes de cuisine. Chaque texte du corpus s'est vu attribué le niveau CECR correspond au niveau du manuel ou du

livre simplifié dont il a été repris.

Pour la seconde étape, nous avons analysé les performances de deux taggeurs sur un échantillon de 100 phrases tirées de notre corpus : TreeTagger (Schmid, 1994), qui est facile d'utilisation et largement reconnu et un taggeur « maison » basé sur des CRF, qui présente l'avantage de détecter un certain nombre d'expressions polylexicales (principalement nominales). Le taggeur CRF s'est révélé légèrement meilleur que le TreeTagger, mais nous avons finalement utilisé ces deux outils en parallèle, car ils comportent chacun des avantages très différents. Nous avons ainsi créé deux versions de FLELex : la première ne comportant pas d'expressions polylexicales. Ces outils ont permis d'annoter l'ensemble des mots du corpus avec leur lemme et leur catégorie de discours. Cette étape de lemmatisation est cruciale, car nous partons du principe que les apprenants de FLE sont sensibles à l'effet de fréquence du lemme plus que celui de la forme fléchie, laquelle répartit la fréquence du lemme parmi tous ses formes fléchies (plus de 50 pour les verbes).

Le dénombrement des lemmes constitue l'étape suivante du processus. Pour ce faire, un script informatique lit le corpus et stocke dans une structure de données la fréquence de chaque lemme dans chaque niveau du CECR. Ces fréquences sont ensuite normalisées à deux niveaux : chaque fréquence est d'abord multipliée par un indice de dispersion avant d'être normalisée pour un corpus d'un million de mots⁸. L'intérêt de la dispersion est de pénaliser les mots à faible fréquence associés à un certain contexte précis (ex. *heaume* associés aux contes de la table ronde) et, au contraire, de donner plus de poids, aux mots qui se retrouvent dans de nombreux textes différents (Carroll *et al.*, 1971). Quant à normaliser chaque fréquence sur un million, cela permet de comparer directement les fréquences de chacun des six niveaux du CECR, ainsi que la fréquence totale de chaque mot dans les six niveaux.

Enfin, une fois la ressource produite, celle-ci est comparée à un dictionnaire numérique, Lexique3 (New *et al.*, 2004) dans le cas de FLELex. Tous les mots qui ne sont pas retrouvés automatiquement dans ce dictionnaire sont vérifiés manuellement. Nous en avons retrouvé 622 pour FLELex. Il s'agit soit d'erreurs de lemmatisation ou de tagging, qui sont corrigées manuellement, soit de mots rares, absents de Lexique3, qui sont alors conservés.

3.4 FLELex : description de la ressource

Au terme de cette procédure méthodologique, nous avons généré deux versions de FLELex : FLELex-TT – basé sur le TreeTagger – comporte 14 236 entrées, tandis que la version basée sur le taggeur CRF, FLELex-CRF, comporte 17 871 entrées⁹. Le Tableau 1 illustre le contenu de la

⁸ Consulter François *et al.* (2014) pour la formule mathématique exacte.

ressource pour quelques entrées. On y voit notamment que certains mots concrets comme *voiture* apparaissent davantage dans les premiers niveaux, tandis que des mots plus abstraits comme *justice* ou *logique* voient leur fréquence augmenter avec les niveaux CECR. L'utilisateur peut profiter de l'interface existant¹⁰ pour faire des recherches sur n'importe quel mot de ces deux bases de données. Il est également possible de comparer les distributions de différents mots ou synonymes, que ce soit en français (ex. *néanmoins* contre *cependant*), comme dans d'autres langues (ex. *maison* vs. *house*).

On note aussi que, malheureusement, eu égard à la taille limitée du corpus, certains mots ou expressions (ex. *sous réserve de*) ne sont observés qu'une seule fois dans le corpus. On parle alors d'hapax. La version FLELex-TT en comporte 33 %, tandis que la version FLELex-CRF comporte 20 % d'hapax. Il convient donc d'être prudent en utilisant ces ressources, puisqu'un nombre non négligeable de mots ont dès lors une distribution qui a été estimée sur un nombre insuffisant de données pour être fiable. Un travail est en cours pour augmenter la taille du corpus et rendre plus robuste les estimations de fréquences.

4 La prédiction des mots inconnus

Le projet CEFRLex présente, en soi, plusieurs caractéristiques intéressantes pour évaluer la difficulté du lexique, mais l'une de ses finalités les plus évidentes est d'évaluer la difficulté lexicale d'un texte au regard du CECR, comme le fait, par exemple pour l'anglais, le site Text inspector¹¹. Pour ce faire, il convient de définir une fonction mathématique capable de transformer la

lemma	tag	A1	A2	B1	B2	C1	C2	total
voiture (1)	NOM	633.3	598.5	482.7	202.7	271.9	25.9	461.5
abandonner (2)	VER	35.5	62.3	104.8	79.8	73.6	28.5	78.2
justice (3)	NOM	3.9	17.3	79.1	13.2	106.3	72.9	48.1
kilo (4)	NOM	40.3	29.9	10.2	0	1.6	0	19.8
logique (5)	NOM	0	0	6.8	18.6	36.3	9.6	9.9
en bas (6)	ADV	34.9	28.5	13	32.8	1.6	0	24
en clair (7)	ADV	0	0	0	0	8.2	19.5	1.2
sous réserve de (8)	PREP	0	0	0.361	0	0	0	0.03

Tableau 1: extrait de FLELex.

⁹ Signalons que cette version n'a pas été nettoyée manuellement, ce qui explique aussi pourquoi il y a davantage d'entrées. Elle est toutefois uniquement utilisée pour ses expressions polylexicales dans l'interface de FLELex en ligne.

¹⁰ <https://cental.uclouvain.be/cefrlex/flelex/search/>

¹¹ <https://textinspector.com/>

distribution de probabilité d'un mot (les fréquences sur les 6 niveaux du CECR) en un niveau CECR unique, qui corresponde au niveau moyen auquel les apprenants de FLE acquièrent ce mot passivement.

Dans le cadre du projet CEFRLex, plusieurs fonctions mathématiques ont été explorées : une première approche, appelée « première occurrence », consiste simplement à attribuer à un mot le premier niveau du CECR dans lequel il a été observé (Gala *et al.*, 2013). Ainsi, à la table 1, le mot *logique* se verrait attribuer le niveau B1. Signalons que cette règle de transformation est la stratégie par défaut pour toute nouvelle ressource du projet CEFRLex. Toutefois, elle n'apparaît guère convaincante pour les praticiens du FLE.

Plus récemment, Alfter *et al.* (2016) ont développé le concept de « significant onset of use » qui consiste à attribuer à un mot le premier niveau CECR dont la fréquence est significativement plus importante que celle du niveau précédent. Todirascu et al. (2019) ont testé une autre règle assez simple de transformation : le mode ou fréquence maximale de la distribution. Enfin, Pintard et François (2020) ont entraîné un algorithme d'apprentissage supervisé de transformation des fréquences de FLELex en un niveau CECR à partir des référentiels de Beacco. Comme il s'agit de la stratégie qui est recommandée pour le français, la section suivante donne davantage de détails sur la conception de ce modèle.

4.1 Transformer les fréquences de FLELex en niveaux CECR avec l'IA

Comme tous les algorithmes statistiques recourant à l'intelligence artificielle, l'objectif de celui-ci est de tirer profit de données afin d'en extraire des motifs de décision qui peuvent être reproduits par l'algorithme sur de nouvelles données. Dans notre cas, les données utilisées sont les mots repris aux chapitres 4 et 6 des référentiels de Beacco (*cf.* section 3.1). Il y a 4 269 mots en commun entre ces référentiels et CEFRLex-TT et nous avons ainsi constitué une ressource pour laquelle on connaît à la fois le niveau CECR proposé par les référentiels de Beacco et la distribution de fréquence dans FLELex. Dès lors, un modèle d'apprentissage automatisé est utilisé pour prédire automatiquement, sur la base des 6 fréquences de niveau (A1 à C2) et de la fréquence totale d'un mot donné, son niveau CECR selon Beacco. L'avantage de ce modèle est qu'il peut être généralisé à l'ensemble des entrées de FLELex.

Le développement de ce modèle, appelé « Beacco » sur l'interface de FLELex, est décrit en détails dans Pintard et François (2020). Pour aller à l'essentiel, disons d'abord que nous avons constitué deux versions des données d'entraînement. Le jeu de données *Beacco.A+B* inclut les 4269 mots des niveaux A1 à B2 en commun entre FLELex et Beacco. Néanmoins, comme il n'y a pas de mots associés au niveau C dans les référentiels de Beacco, nous avons assigné tous les mots de FLELex absents de Beacco à un niveau C global. Une vérification manuelle de cet ensemble a

été réalisée afin de supprimer des mots trop simples, mais absents des référentiels. Enfin, parmi les 9 903 mots niveau C restants, nous avons échantillonné 609 mots au hasard, afin de garder une même proportion de mots par niveau dans notre jeu d'entraînement. Au final, le jeu de données *Beacco A+B+C* inclut 4878 mots.

Dans un second temps, nous avons comparé plusieurs stratégies et algorithmes pour transformer les fréquences de FLELex en un niveau CECR : (1) la méthode de la première occurrence (Gala *et al.*, 2013) ; (2) la méthode du mode (Todirascu *et al.*, 2019) ; (3) une approche basique classique en apprentissage automatisé, qui consiste à attribuer à tous les mots le niveau CECR le plus fréquent dans le corpus d'entraînement (*Classe majoritaire*) ; et trois algorithmes d'apprentissage automatisé, à savoir (4) les arbres de décision ; (5) les machines à vecteurs supports ou SVM et (6) le *boosting*. Les résultats obtenus sur les deux jeux de données sont repris dans le Tableau 2. Nous y présentons une seule mesure d'évaluation : l'exactitude, c'est-à-dire la proportion de mots d'un jeu de test qui sont correctement classés par l'algorithme. On voit que, dans les deux cas, le *boosting* dépasse toutes les autres stratégies. Sur le jeu de données *Beacco A+B+C*, il classe correctement 48 % des mots, faisant ainsi deux fois mieux que les stratégies de base.

Stratégie/algorithme	Exactitude	
	Beacco A+B	Beacco A+B+C
Première occurrence	25 %	27 %
Plus fréquent	18 %	19 %
Classe majoritaire	40 %	22 %
Arbre de décision	52 %	47 %
SVM	53 %	44 %
Boosting	54 %	48 %

Tableau 2 : résultats des différents modèles de transformation de la distribution de fréquence en un niveau CECR unique.

Une dernière expérience réalisée dans Pintard et François (2022) a consisté à entraîner l'algorithme de *boosting* à partir d'une liste dont les fréquences ont été estimées sur des textes pour natifs, à savoir Lexique3. Comme nous l'avons rappelé précédemment, ce type de liste est régulièrement utilisé pour des usages pédagogiques. Néanmoins, l'algorithme perd alors 5 % d'exactitude (sur les données *BeaccoA+B*, il atteint 49 % contre 54 % avec FLELex), ce qui semble confirmer que l'information contenue dans FLELex est plus riche.

Au terme de cette modélisation, nous avons obtenu un modèle d'intelligence artificielle

capable de transformer n'importe quelle distribution de FLELex en un niveau CECR. Ce modèle a été intégré dans l'interface de FLELex et permet, dès lors, d'annoter automatiquement la difficulté de chacun des mots d'un texte, comme illustré à la Figure 2 (figure de droite). La Figure 2 permet également d'illustrer, sur un texte réel, les différences entre la règle de la première occurrence (à gauche) et l'algorithme de *boosting* entraîné sur les référentiels (à droite).

Conclusion

Dans cet article, nous avons montré que le souci d'évaluer les difficultés que pose le lexique à des apprenants de FLE reste vif. Diverses approches ont été mises en œuvre, d'abord manuelles et souvent basées sur des listes de fréquence, puis automatisées et reposant notamment sur le traitement automatique du langage. Aujourd'hui, une première façon d'évaluer la difficulté lexicale d'un texte est de faire appel à une formule de lisibilité, qui produit non seulement une évaluation globale de la difficulté de lecture d'un texte, mais parfois aussi des évaluations plus ciblées sur une dimension du texte, comme sur le site DMesure.

Dans un second temps, nous avons présenté la ressource FLELex, qui s'appuie sur des corpus de textes pédagogiques pour proposer la distribution de fréquence des mots du français sur les 6 niveaux du CECR. Il s'agit d'une autre façon d'aborder la difficulté lexicale pour le FLE, offrant un regard plus fin sur l'usage des mots durant le parcours d'apprentissage en FLE qu'une liste de fréquence classique. Enfin, dans une troisième partie, nous avons expliqué comment il était

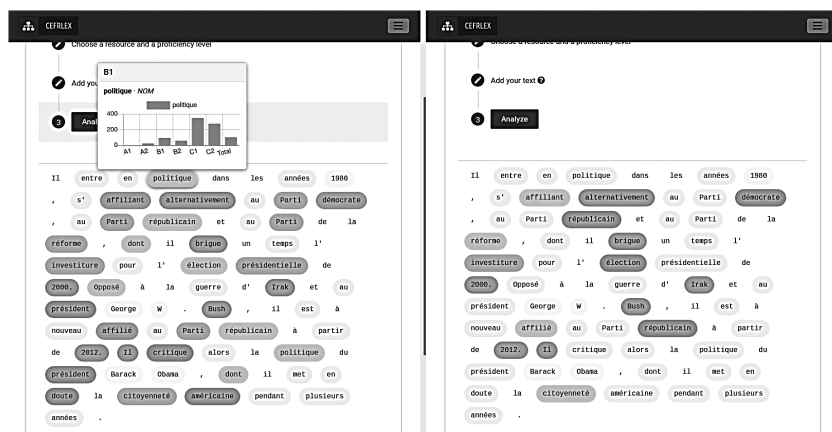


Figure 2 : exemples d'analyse d'un texte à l'aide de l'analyseur de FLELex : à gauche, avec la règle de première occurrence ; à droite, avec le modèle entraîné sur les référentiels de Beacco.

possible de transformer ces distributions de fréquence en un niveau CECR unique en vue d'annoter des textes et d'y identifier les mots difficiles pour des apprenants d'un niveau donné. Pour le futur, il est prévu d'affiner le site Dmesure afin qu'il s'adapte au mieux aux besoins pédagogiques des professeurs de FLE. Nous prévoyons également de publier une nouvelle version de FLELex, basée sur un corpus deux fois plus large.

Remerciements

Nous tenons à remercier Mme Nami Yamaguchi pour sa relecture de certains passages, qui nous a aidé à améliorer la qualité du texte. Nous remercions également tous nos collègues et collaborateurs qui ont participé aux différents projets décrits ici avec beaucoup d'enthousiasme et de compétence.

Références

- Alfter D., Bizzoni Y., Agebjörn A., Volodina E. & Pilán I. (2016). From distributions to labels: a lexical proficiency analysis using learner corpora. In *Proceedings of the NLP4CALL 2016 workshop* (pp. 1-7).
- Brysbart M., Keuleers E. & Mandera P. (2021). Which words do English non-native speakers know? New supranational levels based on yes/no decision. *Second Language Research*, 37(2), 207-231.
- Cha M., Gwon Y. & Kung H. T. (2017). Language modeling by clustering with word embeddings for text readability assessment. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 2003-2006).
- 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.
- Collins-Thompson K. & Callan J. P. (2004). A language modeling approach to predicting reading difficulty. In *Proceedings of HLT-NAACL 2004* (pp. 193-200).
- Dale E. & Chall J. (1948). A formula for predicting readability. *Educational research bulletin*, 27(1): 11-28.
- Devlin J., Chang M. W., Lee K. & Toutanova K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*.
- Filighera A., Steuer T. & Rensing C. (2019). Automatic text difficulty estimation using embeddings and neural networks. In *European Conference on Technology Enhanced Learning* (pp. 335-348). Springer : Cham.
- Flesch R. (1948). A new readability yardstick. *Journal of applied psychology*, 32(3), 221.

- François T., Billami M., Gala N. & Bernhard D. (2016). Bleu, contusion, ecchymose : tri automatique de synonymes en fonction de leur difficulté de lecture et compréhension. In *JEP-TALN-RECITAL 2016. Vol. 2.* (pp. 15-28).
- François T., Gala N., Watrin P. & Fairon C. (2014). FLELex: a graded lexical resource for French foreign learners. In *Proceedings of LREC 2014* (pp. 3766-3773).
- François T. & Miltsakaki E. (2012). Do NLP and machine learning improve traditional readability formulas? In *Proceedings of the PITR workshop* (pp. 49-57).
- François T. & Fairon C. (2012). An "AI readability" formula for French as a foreign language. In *Proceedings of EMNLP 2012* (pp. 466-477).
- François T. (2011). *Les apports du traitement automatique du langage à la lisibilité du français langue étrangère*. Thèse de doctorat, Université catholique de Louvain.
- Gougenheim G., Michéa R., Rivenc P. & Sauvageot A. (1964). *L'élaboration du français fondamental (1er degré)*. Paris : Didier.
- Gunning R. (1952). *The technique of clear writing*. New York : McGraw-Hill.
- Hashimoto B.J. & Egbert J. (2019) More than frequency? Exploring predictors of word difficulty for second language learners. *Language Learning*, 69, 839–72.
- Heilman M., Collins-Thompson K. & Eskenazi M. (2008). An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the BEA 2008 workshop* (pp. 71-79).
- Howes D. & Solomon R. (1951). Visual duration threshold as a function of word probability. *Journal of Experimental Psychology*, 41(40), 1-4.
- Hulstijn J. H. (2007). The shaky ground beneath the CEFR: quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal*, 91(4), 663-667.
- Jurafsky D. & Martin J. H. (2009). *Speech and language processing. 2e édition*. Upper Saddle River (NJ) : Pearson Prentice Hall.
- Kate R., Luo X., Patwardhan S., Franz M., Florian R., Mooney R., Roukos S. & Welty C. (2010). Learning to predict readability using diverse linguistic features. In *Proceedings of Coling 2010* (pp. 546-554).
- Koda K. (2005). *Insights into second language reading: a cross-linguistic approach*. Cambridge : Cambridge University Press.
- Lively B. A. & Pressey S. L. (1923). A method for measuring the vocabulary burden of textbooks. *Educational administration and supervision*, 9(7), 389-398.
- Lonsdale D. & Le Bras Y. (2009). *A frequency dictionary of French: core vocabulary for learners*. New York : Routledge.

- Martinc M., Pollak S. & Robnik-Šikonja M. (2021). Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1), 141-179.
- Nadeem F. & Ostendorf M. (2018). Estimating linguistic complexity for science texts. In *Proceedings of the BEA 2018 workshop* (pp. 45-55).
- New B., Pallier C., Brysbaert M. & Ferrand L. (2004). Lexique 2 : a new French lexical database. *Behavior Research Methods, Instruments, & Computers*, 36(3), 516-524.
- Pintard A. & François T. (2020). Combining expert knowledge with frequency information to infer CEFR levels for words. In *Proceedings of the READI 2020 workshop* (pp. 85-92).
- Schmid H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.
- Schwarm S. E. & Ostendorf M. (2005). Reading level assessment using support vector machines and statistical language models. In *Proceedings of ACL 2005* (pp. 523-530).
- Smith E. & Senter R. (1967). *Automated Readability Index*. Rapport technique, AMRL-TR-66-220, Aerospace Medical Research Laboratories, Wright-Patterson Airforce Base, OH.
- Tack A. (2021) *Mark my words. On the automated prediction of lexical difficulty for foreign language readers*. Thèse de doctorat, KU Leuven & Université catholique de Louvain.
- Tharp J. B. (1939). The measurement of vocabulary difficulty. *The Modern Language Journal*, 24(3), 169-178.
- Thorndike E. (1921). Word knowledge in the elementary school. *The Teachers College Record*, 22(4), 334-370.
- Todirascu A., Cargill M. & François T. (2019). PolylexFLE : une base de données d'expressions polylexicales pour le FLE. In *Actes de TALN 2019* (pp. 143-156).
- Uitdenbogerd S. (2005). Readability of French as a foreign language and its uses. In *Proceedings of the Australian Document Computing Symposium* (pp. 19-25).
- Vander Beke G. E. (1932). *French word book. Vol. 15*. London : Macmillan.
- Vogel M. & Washburne C. (1928). An objective method of determining grade placement of children's reading material. *The Elementary School Journal*, 28(5), 373-381.
- Wilkens R., Alfter D., Wang X., Pintard A., Tack A., Yancey K. & François T. (2022). Fabra: French aggregator-based readability assessment toolkit. In *Proceedings of LREC 2022* (pp. 1217-1233).
- Yancey K., Pintard A. & Francois T. (2021). Investigating readability of French as a foreign language with deep learning and cognitive and pedagogical features. *Lingue e linguaggio*, 20(2), 229-258.