Sylviane GRANGER

CORPUS D'APPRENANTS, ANNOTATION D'ERREURS ET ALAO : UNE SYNERGIE PROMETTEUSE *

Learner corpora, i.e. electronic collections of foreign or second language learner data, constitute a new resource for second language acquisition and foreign language teaching specialists. They are especially useful when they are error-tagged, i.e. when all errors in the corpus have been annotated with the help of a standardized system of error tags. This article describes the three- tiered error annotation system designed to annotate the French Interlanguage Database (FRIDA) corpus. The research took place within the framework of the FreeText project, which aimed to produce a learner corpus-informed CALL program for French as a Foreign Language. Once annotated, the FRIDA corpus was put through standard text retrieval software to extract detailed error statistics and to carry out concordance-based analyses of specific error types. The results were used to focus the CALL exercises on learners' attested difficulties and to improve the error diagnosis system integrated in the CALL program.

1. Corpus d'apprenants

Les corpus d'apprenants, également appelés corpus d'interlangue (IL), sont des recueils électroniques de données authentiques de langue étrangère ou de langue seconde. Ils diffèrent en deux aspects majeurs des types de données généralement utilisées par les chercheurs en acquisition des langues: (a) ils sont *numérisés* et peuvent donc être analysés au moyen d'un large éventail d'outils logiciels permettant une manipulation rapide et efficace des textes et de programmes de traitement automatique du langage qui enrichissent les données d'informations linguistiques (par exemple, la catégorie grammaticale et la structure syntaxique) et (b) ils sont de *grande taille* et constituent donc une base empirique beaucoup plus fiable que celle qui était disponible auparavant pour décrire et modéliser la langue de l'apprenant. La taille est évidemment une notion relative. Un corpus de 200.000 mots est considéré comme grand dans le domaine de l'acquisition des langues où les chercheurs s'appuient habituellement sur des échantillons beaucoup plus petits, mais il apparaît comme minuscule dans le domaine de la linguistique de corpus où le recours à des mégacorpus de plusieurs centaines de millions de mots est devenu la norme plutôt que l'exception.

La langue de l'apprenant diffère de la langue maternelle tant quantitativement que qualitativement. Elle se manifeste par des fréquences de mots, d'expressions et de structures très différentes, certains éléments étant surutilisés et d'autres considérablement sous-utilisés. (Pour des exemples, voir GRANGER, 1998.) Elle est également caractérisée par un taux élevé d'usages impropres, à savoir des erreurs orthographiques, lexicales et grammaticales[1]. Alors que les différences de fréquence peuvent être extraites automatiquement en soumettant un corpus d'apprenants brut, c'est-à-dire non annoté, à un logiciel d'analyse de corpus tel que *WordSmith Tools* (SCOTT, 1996), les erreurs sont plus difficiles à détecter. Les programmes actuels de correction orthographique et grammaticale peuvent traiter un grand nombre

* Cet article est une version traduite et remaniée d'un article publié initialement en anglais : S. GRANGER (2003) : « Error-tagged learner corpora and CALL: a promising synergy », CALICO 20/3, p. 465-480.

[1] Le taux d'erreurs correspond évidemment au niveau de compétence de l'apprenant. Dans nos corpus de français langue étrangère, il varie d'une erreur tous les 5,2 mots à une erreur tous les 9,5 mots, la moyenne étant de 6,89.

d'erreurs d'un locuteur natif mais détectent seulement une minorité d'erreurs propres aux apprenants.

Si on veut pouvoir élaborer des programmes de détection et de correction d'erreurs hautement efficaces, il est nécessaire de récolter au préalable d'importants corpus d'apprenants et d'analyser les erreurs qu'ils contiennent. Dans le cadre du projet FreeText financé par l'Union Européenne[2] visant à produire un programme d'ALAO pour le français langue étrangère (FLE) intégrant des outils de traitement automatique du langage, y compris un système de diagnostic d'erreurs, l'équipe de Louvain a recueilli et étiqueté une importante collection d'écrits en français rédigés par des apprenants de niveau intermédiaire à avancé. Le corpus, appelé *French Interlanguage Database* (*FRIDA*), contient 450.000 mots dont les deux tiers ont été entièrement annotés. Le système d'analyse d'erreurs utilisé pour annoter le corpus est décrit dans la section suivante.

2. Annotation d'erreurs

2.1. *Analyse d'erreurs assistée par ordinateur*

Autrefois très populaire, l'analyse d'erreurs (AE) n'a actuellement plus la cote parmi les chercheurs en acquisition des langues. Elle est passée à la postérité comme une façon d'approcher le langage de l'apprenant confuse, non scientifique et peu fiable. Toutefois, les erreurs font partie intégrante de l'interlangue et sont tout autant dignes d'analyse que tout autre aspect. Ainsi que le souligne RINGBOM (1987 : 69) « Quoique l'analyse d'erreurs ait très certainement ses limites, elle doit être considérée comme une clé importante pour une meilleure compréhension du processus sous-jacent de l'apprentissage L2 »[3]. De même, ELLIS (1994 : 20) note que, quoique les premières études d'AE aient été peu fiables et difficiles à interpréter, « l'étude des erreurs d'apprenants peut encore être un outil utile et est toujours d'actualité »[4]. Plus particulièrement, une description détaillée d'erreurs d'apprenants ne peut que contribuer à un but essentiel de l'apprentissage, à savoir aider les apprenants à atteindre un haut niveau de précision dans la langue cible.

La méthodologie d'AE adoptée dans le projet FreeText est basée sur le système d'analyse d'erreurs assisté par ordinateur développé pour l'anglais par DAGNEAUX, DENNESS et GRANGER (1998). Elle comprend les étapes suivantes :

1. correction manuelle du corpus de FLE ;
2. élaboration d'un système d'étiquetage d'erreurs pour le FLE ;
3. insertion d'étiquettes d'erreurs et de corrections dans les fichiers textes ;
4. extraction de listes de types d'erreurs spécifiques et de statistiques d'erreurs ;
5. analyse linguistique basée sur la concordance des principaux types d'erreurs.

Afin d'être tout à fait efficace, un système d'annotation d'erreurs doit être :

---

[2] Le titre complet du projet est *Le français en contexte : un système hypermédia d'ELAO avancé intégrant des outils de TALN pour un traitement intelligent de documents authentiques et d'exercices de production libre*. Le projet a reçu un support financier de la Commission Européenne dans le programme IST du 5ème Programme-Cadre, contrat IST-1999-13093. Les partenaires du projet étaient l'Université de Genève (Département de Linguistique), l'Université catholique de Louvain (Centre for English Corpus Linguistics), l'Institut de Science et Technologie (Department of Language Engineering) de l'Université de Manchester et la société française Softissimo.

[3] « Although error analysis certainly has its limitations, it must be regarded as an important key to a better understanding of the process underlying L2-learning ».

[4] « The study of learner errors can still serve as a useful tool and is still undertaken ».

1. *informatif* mais *gérable* : il doit être suffisamment détaillé pour fournir des informations utiles sur les erreurs de l'apprenant mais pas au point de devenir ingérable pour l'annotateur ;
2. *réutilisable* : les catégories doivent être suffisamment générales pour être applicables à plusieurs langues ;
3. *souple* : il doit permettre l'ajout ou la suppression d'étiquettes au stade de l'annotation ainsi que leur extraction rapide à l'aide de logiciels d'analyse de corpus ;
4. *cohérent* : pour garantir un maximum de cohérence entre annotateurs, une description détaillée des catégories d'erreurs et des principes d'étiquetage d'erreurs doit être incluse dans un manuel.

Nous avons pris soin de nous assurer que ces conditions étaient remplies au moment de la conception du système d'étiquetage d'erreurs pour le FLE.

2.2. *Système d'étiquetage d'erreurs*

DULAY, BURT et KRASHEN (1982, ch. 7) proposent deux taxonomies principales de description d'erreurs : (a) l'une basée sur les catégories linguistiques (soit générales telles que morphologie, lexique et grammaire, soit plus spécifiques telles qu'auxiliaires, passifs et prépositions), (b) l'autre axée sur la manière dont les structures de surface ont été altérées par les apprenants (par exemple, omission, addition ou ordre erroné). Comme JAMES (1998), nous estimons qu'il y a un grand avantage à combiner ces deux approches en une seule taxonomie bidimensionnelle, voire même, moyennant l'ajout d'une couche supplémentaire d'information sur les erreurs, en une taxonomie tridimensionnelle.

Le système d'étiquetage des erreurs développé pour annoter le corpus *FRIDA* se compose de trois niveaux d'annotation : le domaine d'erreurs, la catégorie d'erreurs et la catégorie de mots. Ces trois niveaux sont descriptifs plus qu'interprétatifs. Nous avons délibérément décidé de ne pas utiliser de distinctions telles que « erreurs » et « fautes » ou erreurs interlinguales et intralinguales qui sont difficiles à attribuer et qu'il vaut mieux garder pour une étape ultérieure de l'analyse.

2.2.1. *Domaine et catégorie d'erreurs*

Le domaine d'erreurs représente le niveau le plus général : il précise si l'erreur est formelle (c.-à-d. orthographique), grammaticale, lexicale, etc. Chaque domaine d'erreurs est subdivisé en un nombre variable de catégories d'erreurs. Le Tableau 1 donne le détail des neuf domaines d'erreurs.

Tableau 1 : Domaines et catégories d'erreurs

Domaines d'erreurs		Catégories d'erreurs	
<F>	Forme	<AGL>	Agglutination
		<MAJ>	Majuscule/minuscule
		<DIA>	Signe diacritique
		<HOM>	Homonymie
		<GRA>	Autres erreurs d'orthographe
<M>	Morphologie	<MDP>	Dérivation-préfixation
		<MDS>	Dérivation-suffixation
		<MFL>	Inflexion
		<MFC>	Inflexion-confusion
		<MCO>	Composition

<G>	Grammaire	<CLA>	Classe
		<AUX>	Auxiliaire
		<GEN>	Genre
		<MOD>	Mode
		<NBR>	Nombre
		<PER>	Personne
		<TPS>	Temps
		<VOI>	Voix
		<EUF>	Euphonie
<L>	Lexique	<SIG>	Signification
		<CPA>	Complémentation adjectivale
		<CPD>	Complémentation adverbiale
		<CPV>	Complémentation verbale
		<CPN>	Complémentation nominale
		<FIG>	Figement
<X>	Syntaxe	<ORD>	Ordre des mots
		<MAN>	Mot manquant
		<RED>	Mot redondant
		<COH>	Cohésion
<R>	Registre	<RLE>	Lexique
		<RSY>	Syntaxe
<Y>	Style	<CLR>	Obscur
		<LOU>	Lourd
<Q>	Ponctuation	<CON>	Ponctuation erronée
		<TRO>	Ponctuation redontante
		<OUB>	Ponctuation manquante
<Z>	Coquille		

Vu les contraintes de place, nous nous contentons de donner une brève description d'un seul domaine, celui des erreurs lexicales. Les autres catégories sont illustrées à l'aide d'un exemple représentatif dans l'annexe A.

Le domaine lexical <L> regroupe toutes les erreurs lexicales dues à :

1. une connaissance insuffisante du sens conceptuel des mots : <SIG> (voir exemple 1) ;
2. des violations des modes de co-occurrence des mots. Cette catégorie couvre un large spectre allant des collocations restreintes aux idiomes : <FIG> (voir exemple 2) ; et
3. des violations des modes de complémentation grammaticale des mots. Cette catégorie comprend la valence des verbes <CPV>, des noms <CPN>, des adjectifs <CPA> et des adverbes <CPD> (voir exemples 3 et 4).

(1) *Il est un peu plus* ****commis à*** *[engagé envers] l'idéal européen*. <SIG>
(2) *Leurs élèves suivent des* ****cours religieux*** *[cours de religion]*. <FIG>
(3) *…bien qu'il soit libre* ****à*** *[de] choisir ce qu'il fait.* <CPA>
(4) *Le Canada pourrait se transformer* ****dans*** *[en] un type de superstructure*. <CPV>

2.2.2. *Catégorie de mots*

La catégorie de mots de l'élément erroné est étiquetée au moyen d'un système d'étiquetage morphosyntaxique comprenant 11 catégories principales subdivisées en 54 sous-

catégories. (Voir l'annexe B pour une liste complète des étiquettes). L'ajout au système de ce troisième niveau rend possible le tri d'erreurs par catégorie grammaticale et l'établissement d'une liste pertinente de catégories d'erreurs pour chacune d'elles. Il y a, par exemple, quatre catégories d'erreurs pertinentes différentes pour les erreurs de prépositions : deux catégories lexicales (erreurs sémantiques et de complémentation) et deux syntaxiques (prépositions manquantes et redondantes). Nous n'avons pas utilisé de programmes d'étiquetage automatique ne sachant pas dans quelle mesure les erreurs des apprenants pourraient affecter le taux de réussite des programmes.

2.2.3. *Correction*

Les formes correctes sont également insérées dans les fichiers de textes en regard des formes erronées : (a) pour faciliter l'interprétation ultérieure des annotations d'erreurs et, (b) pour permettre le tri automatique au niveau des formes correctes. La Figure 1 contient un extrait de la concordance des erreurs du type <F><GRA><NOM>. Un tri sur la forme correcte *développement* fait clairement ressortir les sept variantes erronées attestées dans le corpus.

Figure 1 : Concordance de <F><GRA><NOM> : Tri sur les corrections

(1)	tourné vers le	*<F><GRA><NOM># **développement$*** **dévélopement** culturel
(2)	Les résultats du	*<F><GRA><NOM>#**développement$*** **dévélopment** technique
(3)	On a intensifié le	*<F><GRA><NOM># **développement$*** **développement** des sports.
(4)	Les chantiers de	*<F><GRA><NOM>#**développement$*** **Dévelopement**.
(5)	On est pour le	*<F><GRA><NOM>#**développement$*** **dévelopment** de l'avenir.
(6)	Elle connaît un bon	*<F><GRA><NOM>#**développement$*** **developement**.
(7)	La conséquence de ce	*<F><GRA><NOM>#**développement$*** **developpment**, limité

Une mise en garde s'impose. Alors que la correction est incontestable dans certains cas (voir exemples 2-4 ci-dessus), dans d'autres (cf. exemple 1), elle n'est qu'une possibilité parmi plusieurs autres que l'annotateur aurait pu apporter. En utilisant la base de données, il est donc important de garder à l'esprit que certaines corrections n'ont qu'une valeur indicative.

2.2.4. *Insertion d'étiquettes*

Pour accélérer le processus d'insertion d'étiquettes, nous avons développé un éditeur de texte piloté par menu qui permet à l'annotateur d'insérer dans le texte des étiquettes d'erreurs à l'endroit voulu en cliquant sur l'étiquette appropriée du menu. En utilisant la boîte de correction, l'analyste peut également insérer la forme corrigée avec les symboles de formatage appropriés.

Figure 2 : Exemple de texte annoté

L'héritage du passé est très <G><GEN><ADJ> #fort$ forte </ADJ></GEN></G> et le sexisme est toujours présent. Beaucoup de gens pensent que la femme est un être pas très intelligent, qui bavarde beaucoup et qui ne sait pas garder le moindre <F><DIA><NOM> #secret$ secrèt </NOM></DIA></F>. Ces gens <G><NBR><VSC> #pensent$ pense </VSC></NBR></G> aussi que les femmes ne sont pas aptes à prendre des responsabilités.

Dans l'exemple de texte annoté de la Figure 2, trois erreurs ont été annotées : deux erreurs grammaticales – erreur d'accord de genre de l'adjectif *fort* et erreur d'accord de nombre du verbe *penser* – et une erreur formelle, le signe diacritique incorrect du nom *secret*.

3. Statistiques et analyse d'erreurs

Dans le cadre du projet FreeText, une grande partie du corpus FRIDA (300.000 mots) a été entièrement annotée par l'équipe de Louvain. Un total de 46.241 erreurs a été détecté manuellement et les étiquettes et corrections appropriées insérées dans les fichiers textes au moyen de l'éditeur d'erreurs.

Un des principaux avantages du système d'étiquetage à trois niveaux est qu'il permet une large gamme de recherches, de la plus générale à la plus spécifique. Il est possible d'extraire toutes les erreurs dans un domaine particulier (p.ex. <Q>, soit toutes les erreurs de ponctuation) ou une catégorie particulière (p.ex. <GEN>, soit toutes les erreurs de genre). Il est également possible de faire une recherche sur une catégorie spécifique de mots (par exemple, une recherche sur <ADJ> détectera toutes les erreurs qui affectent les adjectifs). La requête la plus spécifique implique des trigrammes complets. Ainsi, une recherche sur <G><NBR><VSP> générera une liste de toutes les formes participiales des verbes contenant une erreur de nombre.

Les statistiques détaillées extraites du corpus permettent de classer les domaines et les catégories d'erreurs en ordre décroissant de fréquence. Comme indiqué dans le Tableau 2, deux domaines – grammaire et forme – représentent 50 % de toutes les erreurs dans le corpus.

Tableau 2 : Répartition des domaines d'erreurs

Etiquette	Nombre d'occurrences	Pourcentage
G	11.779	25,38
F	11.452	24,68
L	7.198	15,51
X	7.061	15,21
Q	5.707	12,29
R	1.402	3,02
Y	862	1,85
M	784	1,68
Z	155	0,33

Le Tableau 3 donne la liste des 20 premiers trigrammes d'erreurs et montre que l'erreur la plus fréquente dans tout le corpus appartient au domaine de la ponctuation <Q>. A elle seule, le type d'erreur 'virgule manquante' <OUB><PUV> compte pour 8,8 % du nombre total d'erreurs dans le corpus. Il est intéressant de noter que les 10 premiers trigrammes représentent environ un tiers (31 %) des erreurs dans le corpus, les 20 premiers 45 % et les 50 premiers 70 % des erreurs[5].

Une fois insérés dans les fichiers textes, les codes d'erreurs peuvent être recherchés en utilisant un outil tel que *WordSmith Tools*. Les concordances d'étiquettes d'erreurs permettent à l'analyste de visualiser les erreurs dans leur contexte et de les trier de plusieurs manières. La Figure 3 montre la concordance des erreurs de complémentation adjectivale <CPA> impliquant des prépositions <PES>. Ce processus fait ressortir quelques adjectifs particulièrement sujets aux erreurs tels que *différent*, *difficile* et *nécessaire*.

[5] Il est également intéressant de noter que ces 50 trigrammes représentent seulement 9 % du nombre total des différents trigrammes d'erreurs attestés dans le corpus (N = 567).

Tableau 3 : Les 20 premiers trigrammes d'erreurs

	Trigramme d'erreur			Nombre d'occurrences	Pourcentage cumulatif
1	Q	OUB	PUV	4.805	8,80
2	F	GRA	NOM	1.610	12,26
3	L	SIG	PES	1.592	15,69
4	F	DIA	NOM	1.485	18,88
5	G	GEN	ADJ	1.306	21,69
6	G	TPS	VSC	1.200	24,27
7	L	FIG	SEQ	914	26,23
8	X	MAN	ADE	812	27,97
9	F	MAJ	NOM	783	29,65
10	G	NBR	NOM	765	31,29
11	L	SIG	VSC	745	32,89
12	L	CPV	PES	710	34,41
13	L	SIG	NOM	708	35,93
14	F	DIA	ADJ	689	37,41
15	X	MAN	PES	619	38,74
16	F	GRA	ADJ	609	40,05
17	R	RSY	SEQ	584	41,03
18	G	NBR	ADJ	582	42,55
19	G	CLA	AIN	580	43,79
20	G	MOD	VSC	523	44,91

Figure 3 : Concordance de <L><CPA><PES>

selon la nature, tout à fait *différent*	<L><CPA><PES>***#de$*** **à** celui que
expériences personnelles *différentes*	<L><CPA><PES>***#de$*** **à** celles de
beaucoup plus *différente*	<L><CPA><PES>***#de$*** **que** celle
dont les priorités sont *différentes*	<L><CPA><PES>***#de$*** **que** celles des autres
on peut dire qu'il est bien *difficile*	<L><CPA><PES>***#de$*** **à** résister
il devient très *difficile*	<L><CPA><PES>***#de$*** **à** combattre
Il était *difficile*	<L><CPA><PES>***#de$*** **à** savoir
Il est *nécessaire*	<L><CPA><PES>***#de$*** **0** connaître l'anglais[6]
Il est *nécessaire*	<L><CPA><PES>***#d'$*** **0** avoir une bonne santé

4. Intégration des résultats dans des outils d'ALAO

Dans le cadre du projet FreeText, des statistiques et analyses d'erreurs furent utilisées : (a) pour sélectionner les domaines linguistiques sur lesquels se concentrer dans le programme d'ALAO et pour adapter les exercices en fonction des types d'erreurs attestés, et, (b) pour convertir un correcteur d'orthographe du français existant[7] et un analyseur syntaxique du français[8] en un système intégré de diagnostic d'erreurs qui vérifie les réponses des apprenants et fournit d'utiles retours d'information.

Les catégories les plus sujettes aux erreurs ont fait l'objet d'une attention particulière dans les exercices d'ALAO : erreurs de temps et de mode, d'accord (nombre et genre), d'articles, de complémentation et erreurs de prépositions. Les exercices ont été conçus afin de

[6] The symbol '0' is used to represent missing elements.
[7] Le correcteur d'orthographe est la composante vérificateur orthographique du logiciel Softissimo *Hugo 2000.*
[8] L'analyseur syntaxique est l'analyseur FIPS de l'Université de Genève (voir WEHRLI, 1997).

refléter le type de contexte dans lequel les apprenants commettent le plus d'erreurs. Par exemple, pour l'accord du participe passé, on a pris soin d'inclure une grande proportion de sujets féminins étant donné que l'accord de genre se révèle être particulièrement problématique (*Pour cela la chambre à coucher et aussi la literie doivent être soigneusement aérées*) ainsi que de longs groupes sujets dont le noyau est relativement éloigné du verbe (*Les couvertures de laine et les édredons qu'on utilise pour recouvrir certains meubles doivent toujours être aérés chaque semaine*)[9]. Les erreurs formelles, représentant un quart des erreurs dans le corpus, sont visées au travers de dictées et d'une série d'exercices visant des difficultés spécifiques : homonymes ainsi que mots et phrases portant souvent à confusion (p.ex. *peut-être* et *peut être*), termes de nationalité (*parler l'anglais* ; *les Anglais*), etc. Enfin, alors que nos statistiques d'erreurs montrent qu'elle constitue une difficulté majeure pour les apprenants, la ponctuation, parent pauvre de la plupart des programmes d'ALAO, est dûment prise en compte.

Les statistiques et analyses d'erreurs ont également été utilisées pour adapter des outils de TALN existants (correcteur d'orthographe et analyseur syntaxique) et les transformer en un outil de diagnostic performant. Le système qui en résulte peut traiter les erreurs orthographiques et bon nombre d'erreurs grammaticales (accords de nombre et de genre, euphonie, etc.), certaines erreurs syntaxiques (ordre des adjectifs et adverbes) et quelques erreurs lexicales (complémentation) mais il est impuissant face à un grand nombre d'erreurs qui dépassent les capacités des analyseurs syntaxiques, telles que la plupart des erreurs de ponctuation, des erreurs lexicales sémantiques ou des erreurs de temps. De l'avis même d'un des auteurs, « les performances du système de diagnostic peuvent sembler relativement faibles par rapport aux espoirs soulevés » (L'HAIRE, 2004). Ceci est notamment dû au trop grand nombre de surdétections, particulièrement gênantes pour des utilisateurs dont le français n'est pas la langue maternelle. Néanmoins, le projet a démontré l'intérêt d'utiliser de larges corpus d'apprenants annotés pour affiner les systèmes de diagnostic et identifier de manière plus précise les types d'erreurs qui sont à la portée des techniques d'analyse actuelles. En outre, le corpus FRIDA permet une multitude d'applications pédagogiques. Il est notamment l'élément central d'un « exemplier » (forme de concordancier web), destiné à la fois aux enseignants et aux chercheurs, qui permet d'interroger le corpus et d'en extraire des phrases comportant certains types d'erreurs (pour une description de l'interface, voir GRANGER *et al.*, 2007). Un outil similaire développé pour l'anglais langue étrangère est utilisé directement par les apprenants dans des activités d'éveil au langage (HEGELHEIMER, 2006 ; HEGELHEIMER et FISHER, 2006).

5. Conclusion

L'annotation d'erreurs de corpus d'apprenants est une tâche très minutieuse qui demande beaucoup de temps. Toutefois, une fois le corpus annoté, le retour sur investissement est énorme.

Le système d'annotation d'erreurs à trois niveaux conçu pour le projet a démontré une grande efficacité. Il contient un nombre limité de catégories par niveau (9 pour les domaines, 36 pour les catégories et 54 pour les catégories de mots), ce qui facilite le travail de l'annotateur et permet, en même temps, d'effectuer des analyses relativement précises grâce au grand nombre de combinaisons possibles (567 trigrammes d'erreurs attestés dans le corpus). Un autre avantage du système est sa flexibilité dans la manipulation automatique des données. En utilisant un outil de traitement de corpus tel que *WordSmith Tools*, il est possible d'extraire un des trois niveaux d'annotation, soit séparément soit conjointement, et de trier les

[9] Pour une analyse plus détaillée des erreurs de participes, voir GRANGER, VANDEVENTER et HAMEL (2001).

lignes de concordance de diverses façons afin d'identifier les structures d'erreurs récurrentes. La décision d'insérer des corrections dans les fichiers textes s'est avérée positive, car elle permet de se concentrer sur un élément linguistique spécifique (p.ex. *que*) et d'extraire toutes les formes incorrectes qu'il peut générer (forme zéro – *ils trouvent *0 l'école est une expérience difficile, qui – des tâches *qui ne font pas les Français,* et *comme – pas aussi dur *comme chez eux*).

En limitant le nombre d'étiquettes et en fournissant aux annotateurs un manuel d'étiquetage d'erreurs complet, il a été possible de systématiser en grande partie l'annotation d'erreurs. Il est cependant important de se rendre compte que l'annotation d'erreurs contiendra toujours une part de subjectivité, la notion d'erreur étant particulièrement floue. Comme le font très justement remarquer MILTON et CHOWDHURY (1994 : 129), l'annotation d'erreurs nous permet tout au plus de systématiser nos intuitions. A cet égard, elle est intrinsèquement différente d'autres types d'annotations moins floues telles que l'annotation morphosyntaxique.

Il est également important de garder à l'esprit que, malgré ses nombreux avantages, l'étiquetage d'erreurs se préoccupe uniquement du mauvais usage par l'apprenant. Il n'est pas capable d'identifier d'autres aspects de l'interlangue tels que le suremploi ou le sous-emploi de mots et de phrases qui, associés aux erreurs pures et simples, trahissent l'origine allophone des apprenants. Il n'est pas nécessaire de disposer d'un corpus d'apprenant annoté pour accéder à ces différences de fréquence. Une comparaison entre un corpus d'apprenant brut, sans annotation, et un corpus comparable d'un locuteur natif révèlera automatiquement les mots et les phrases suremployées ou sous-employées de manière significative par les apprenants. Par manque de temps, nous n'avons pu utiliser cette approche que sporadiquement dans le projet FreeText encore que suffisamment pour nous convaincre qu'elle constitue un complément idéal à l'annotation d'erreurs. Par exemple, en ce qui concerne l'usage du pronom, la fonction de comparaison automatique de listes de fréquences de *WordSmith Tools*[10] a permis de mettre en évidence un suremploi du pronom *y* par les apprenants. Une analyse des lignes de concordances montre que ce suremploi est dû à un suremploi massif de la structure impersonnelle *il y a* (*il y a beaucoup de tensions*), couplé à un sous-emploi manifeste d'utilisations anaphoriques de *y* (*la Belgique refuse d'y prendre part*).

Utilisé avec discernement et en pleine connaissance des limitations mentionnées ci-dessus, un corpus doté d'annotations d'erreurs se révèle être un outil précieux pour améliorer notre connaissance de l'interlangue et pour adapter en conséquence le matériel pédagogique, notamment les programmes d'ALAO. Les statistiques détaillées rendent possible l'identification des catégories d'erreurs les plus fréquentes tandis que les concordances de types spécifiques d'erreurs permettent aux analystes de voir les erreurs dans leur contexte et de produire des descriptions fiables de l'interlangue. Bien qu'elle ne constitue certainement pas le but suprême des recherches sur l'interlangue, l'annotation d'erreurs contribue de manière précieuse à mettre en lumière les difficultés des apprenants, tant quantitativement que qualitativement. A ce titre, elle ne peut qu'être bénéfique pour tous les outils pédagogiques d'apprentissage des langues étrangères et, en particulier, les programmes d'ALAO.

Sylviane GRANGER
Centre for English Corpus Linguistics
Université catholique de Louvain

[10] Cette fonction compare automatiquement tous les mots dans des listes de fréquence de deux corpus différents et fait ressortir tous ceux qui apparaissent significativement plus souvent dans l'un que dans l'autre. Pour cette comparaison, nous avons utilisé le corpus FRIDA et un corpus de dissertations écrites par des étudiants francophones de licence en langues romanes à Louvain.

BIBLIOGRAPHIE

DAGNEAUX, E., DENNESS, S. & GRANGER, S. (1998) : « Computer-aided Error Analysis », *System. An International Journal of Educational Technology and Applied Linguistics* 26 (2), p. 163-174.

DULAY, H., BURT, M. & KRASHEN, S. (1982) : *Language Two.* New York: Oxford University Press.

ELLIS, R. (1994) : *The Study of Second Language Acquisition*. Oxford: Oxford University Press.

GRANGER, S. (éd.) (1998) : *Learner English on Computer*. London & New York : Addison Wesley Longman.

GRANGER, S., KRAIF, O., PONTON, C., ANTONIADIS, G. & ZAMPA, V. (2007) : « Integrating learner corpora and natural language processing: A crucial step towards reconciling technological sophistication and pedagogical effectiveness », *ReCALL* 19(3), p. 252-268.

GRANGER, S., VANDEVENTER, A., HAMEL, M.-J. (2001) : « Analyse des corpus d'apprenants pour l'ELAO basé sur le TAL », *Linguistique de corpus*. Numéro spécial de *Traitement automatique des langues* (Hermès, Paris), Vol. 42(2), p. 609-621.

HEGELHEIMER, V. (2006) : « Helping ESL writers through multimodal, corpus-based, online grammar resources », *CALICO Journal* 24(1), p. 1-28.

HEGELHEIMER, V. et FISHER, D. (2006) : « Grammar, writing, and technology: A sample technology-supported approach to teaching grammar and improving writing for ESL learners », *CALICO Journal* 23(2) : p. 257-279.

JAMES, C. (1998) : *Errors in Language Learning and Use. Exploring Error Analysis*. London & New York : Longman.

L'HAIRE, S. (2004) : « Vers un feedback plus intelligent. Les enseignements du projet *Freetext* », *Actes de la Journée d'étude de l'ATALA*, Grenoble, 22 octobre 2004. http://w3.u-grenoble3.fr/lidilem/talal/

MILTON, J. et CHOWDHURY, N. (1994) : « Tagging the interlanguage of Chinese learners of English », in L. FLOWERDEW & A.K. Tong (eds.) *Entering Text.* Hong Kong: The Hong Kong University of Science and Technology, p.127-143.

RINGBOM, H. (1987) : *The Role of the First Language in Foreign Language Learning*. Clevedon & Philadelphia: Multilingual Matters.

SCOTT, M. (1996) : *WordSmith Tools.* Oxford: Oxford University Press.

WEHRLI, E. (1997) : *L'analyse syntaxique des langues naturelles : Problèmes et méthodes.* Paris : Masson.

Annexe A : Domaines et catégories d'erreurs : exemples authentiques

<F> **<AGL>** le **portebagages** (porte-bagages) est sur le toit
<MAJ> Mme Thatcher incarnait la peur des **anglais** (Anglais)
<DIA> Il existe une **ambiguité** (ambiguïté)
<HOM> Ils **ce** (se) déshabillent
<GRA> un **labirint** (labyrinthe)

<M> **<MDP>** le sentiment de **malcontentement** (mécontentement)
<MDS> ...qui continue à **évolutionner** (évoluer)
<MFL> Elle s'occupe des enfants et des **travails** (travaux)
<MFC> Il n'a pas très bien **réussit** (réussi)
<MCO> la participation de **celles-dernières** (celles-ci)

<G> **<CLA>** L'unique chose **que** (qui) n'est pas bonne en Belgique...
<AUX> Je m'**avais** (étais) très bien amusé
<GEN> La protection sociale a été **amélioré** (améliorée)
<MOD> Bien qu'ils **sont** (soient) pressés, ...
<NBR> Elle reprit ses **esprit** (esprits)
<PER> J'espère **s'** (m') adapter rapidement
<TPS> Un sondage qui **était** (a été) publié dans le Monde montre que...
<VOI> Les éclipses **ont vues** (sont vues) comme des présages
<EUF> En prenant **ce** (cet) aspect, ...

<X> **<ORD>** Je peux **m'amuser bien** (bien m'amuser).
<MAN> Je crois **0** (qu') ici il y a beaucoup plus de soirées
<RED> le domaine social **et** (,) économique et politique
<COH> A la métropole, il existe plus d'allocations et d'aide pour les chômeurs, les handicapés et les personnes âgées, **et** (tandis que) dans les îles, il n'y a pas beaucoup de soutien.

<R> **<RLE>** **Quand même** (néanmoins), c'est une histoire lointaine
<RSY> Etant donné que **j'ai pas** (je n'ai pas) souvent présenté mes travaux...

<Y> **<CLR>** Quand la famille d'un travailleur étranger vient le rejoindre, **on** (??) est obligé d'organiser toutes les formalités avant de quitter son pays.
<LOU> mais il y a des autres choses qui **m'ont donné une très grande surprise** (qui m'ont fort surpris)

<Q> **<CON>** La langue devient plus française **-** (:) on l'appelle maintenant le créole francisé.
<TRO> Ce sont surtout les différences biologiques**,** (0) qui sont présentées dans la société avec la grande force de l'organisme humaine.
<OUB> Quand j'ai du temps libre **0** (,) je veux faire des choses reposantes...

<Z> **qunad** (quand); **ps** (pas)

Annexe B : Catégories grammaticales

CATEGORIE GRAMMATICALE		ETIQUETTE
ADJECTIF	Simple	ADJ
	Comparatif	AJC
	Superlatif	AJX
	Complexe	AJL
ADVERBE	Simple	ADV
	Complexe	AVL
ARTICLE	Défini	ADE
	Indéfini	AIN
	Partitif	APA
	Contracté	ACO
CONJONCTION	Coordination	COC
	Subordination simple	COS
	Subordination	COL
DETERMINANT	Démonstratif	DED
	Possessif	DEP
	Indéfini	DEI
	Exclamatif-if	DEX
	Relatif	DER
	Numéral	DEN
NOM	Commun Simple	NOM
	Commun Composé	NOC
	Commun Complexe	NOL
	Propre	NOP
PREPOSITION	Simple	PES
	Complexe	PEL

CATEGORIE GRAMMATICALE		ETIQUETTE
PRONOM	Démonstratif	POD
	Possessif	POP
	Personnel	POO
	Indéfini	POI
	Exclamatif-Interrogatif	POX
	Numéral	PON
	Adverbial	POA
	Relatif	POR
	Impersonnel	POS
VERBE	Fini simple	VSC
	Participe simple	VSP
	Gérondif simple	VSG
	Infinitif simple	VSI
	Fini complexe	VCC
	Participe complexe	VCP
	Gérondif complexe	VCG
	Infinitif complexe	VCI
PONCTUATION	Point	PUP
	Point d'interrogation	PUI
	Point d'exclamation	PUE
	Virgule	PUV
	Point-virgule	PUG
	Deux-points	PUD
	Points de suspension	PUS
	Parenthèses	PUA
	Crochets	PUC
	Guillemets	PUL
	Tiret	PUT
	Barre oblique	PUO
SEQUENCE		SEQ