TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Quel apport des unités polylexicales dans une formule de lisibilité pour le français langue étrangère

Thomas François[1, 2, 3] Patrick Watrin[2, 3]
(1) Aspirant FNRS
(2) Centre de traitement automatique du langage (CENTAL), UCLouvain
(3) Institut Langage et Communication (IL&C), UCLouvain
thomas.francois@uclouvain.be, patrick.watrin@uclouvain.be

Résumé. Cette étude envisage l'emploi des unités polylexicales (UPs) comme prédicteurs dans une formule de lisibilité pour le français langue étrangère. À l'aide d'un extracteur d'UPs combinant une approche statistique à un filtre linguistique, nous définissons six variables qui prennent en compte la densité et la probabilité des UPs nominales, mais aussi leur structure interne. Nos expérimentations concluent à un faible pouvoir prédictif de ces six variables et révèlent qu'une simple approche basée sur la probabilité moyenne des n-grammes des textes est plus efficace.

Abstract. This study considers the use of multi-words expressions (MWEs) as predictors for a readability formula for French as a foreign language. Using a MWEs extractor combining a statistical approach with a linguistic filter, we define six variables. These take into account the density and the probability of MWEs, but also their internal structure. Our experiments conclude that the predictive power of these six variables is low. Moreover, we show that a simple approach based on the average probability of n-grams is a more effective predictor.

Mots-clés : Lisibilité du FLE, unités polylexicales nominales, modèles N-grammes.

Keywords: Readability of FFL, nominal MWEs, N-grams models.

1 Introduction

Avec le succès des approches communicatives et actionnelles dans l'enseignement des langues étrangères, les professeurs sont aujourd'hui encouragés à travailler avec des textes authentiques, afin de mettre leurs étudiants en contact avec des données linguistiques avérées. Le web constitue une précieuse source pour ce type de documents, mais la recherche d'un document adapté au niveau des étudiants reste parfois ardue. Dans ce contexte, la lisibilité a un rôle à jouer, car elle vise à développer des outils capables de prédire les difficultés de compréhension d'un texte pour une population donnée, uniquement au moyen de caractéristiques textuelles (telles que le nombre de lettres par mots, le nombre de mots par phrase, etc.).

Cependant, si de nombreux travaux se sont penchés sur la lisibilité de l'anglais L1, on compte nettement moins d'études sur la lisibilité en L2, et moins encore pour le français langue étrangère (FLE). Dans la majorité des cas, ce sont des formules développées sur des natifs qui ont été appliquées à des contextes d'enseignement des L2. Or, la validité d'une telle approche est loin d'être établie, car elle repose sur trois hypothèses fragiles : (1) la compréhension des lecteurs en L2 est comparable à celle de natifs ; (2) les variables textuelles considérées dans ces formules sont pertinentes pour la lecture en L2 et (3) la pondération de ces variables peut être la même au sein d'une formule pour une L1 et une L2.

Plusieurs auteurs (Laroche, 1979; Uitdenbogerd, 2005) ne s'accordent pas avec ce point de vue et considèrent que les particularités du processus de lecture en L2, décrits entre autres par Koda (2005), doivent être pris en compte dans les formules de lisibilité spécifiques aux L2. Nous avons précédemment montré que les temps et modes verbaux constituent d'excellents prédicteurs pour la lisibilité des textes en FLE (François, 2009). Dans cette étude, nous envisageons une autre facette textuelle a priori intéressante : les unités polylexicales (UPs) nominales. En effet, des expérimentations que nous avons réalisées précédemment ont confirmé que la fréquence des mots de contenu et, en particulier des noms, apportent une information utile en lisibilité. Par ailleurs, les UPs sont associées à une pratique fluide et appropriée de la langue (Pawley & Syder, 1983). On peut donc s'attendre à ce que les apprenants d'une L2, et en particulier les débutants, rencontrent des difficultés à les traiter.

Dans la section 2, nous synthétisons un ensemble de résultats de recherche à propos des UPs et de leur traitement, en particulier lors de la lecture d'un texte en L2. La section 3 décrit la procédure expérimentale qui a été appliquée pour analyser la relation entre les UPs et la difficulté des textes pour des lecteurs de FLE. Les résultats de ces expérimentations sont rapportés et discutés dans la section 4.

2 Les unités polylexicales et la difficulté des textes

Dans cet article, nous désignons sous le terme « unité polylexicale » un ensemble d'objets linguistiques dont le sens et la forme peuvent être plus ou moins figés (collocations, mots composés, expressions idiomatiques, etc.). D'un point de vue statistique, cette classe d'objets réfère communément à 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).

Plusieurs travaux en psychologie cognitive de la lecture, dont celui d'Underwood *et al.* (2004), ont observé qu'en moyenne, le traitement des unités polylexicales par des natifs est plus rapide que celui des chaînes libres, à la lecture comme lors de la production orale. Un effet similaire a aussi été observé pour des apprenants de niveau avancé (Underwood *et al.*, 2004). Cependant, ces études portent sur le temps de lecture et, donc, de reconnaissance des UPs, mais elles n'évaluent pas leur effet sur la compréhension. Or, chez des apprenants débutants ou intermédiaires, il y a fort à parier que cet effet soit contrebalancé par le fait que les unités polylexicales 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.

En lisibilité, l'hypothèse qu'une UP plus rare dans la langue serait plus ardue à la lecture n'a guère été explorée. Weir & Anagnostou (2008) ont suggéré de se servir de la moyenne des fréquences absolues des collocations présentes dans un texte comme d'un indice de sa difficulté, sans valider leur approche par des expérimentations. 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 n'apparaît cependant pas significatif au sein de leur modèle de régression linéaire multiple, puisqu'il

obtient un t [1] de $0,4987$, ce qui correspond à une p-valeur de $0,6188$ (Ozasa *et al.*, 2007, 4).

Ainsi, il n'est pas évident que les UPs constituent un prédicteur efficace en lisibilité en L2. Comme les recherches précitées n'ont abordé cette problématique que superficiellement, nous avons voulu l'explorer plus en détail au travers du cas particulier du FLE.

3 Procédure expérimentale

Pour réaliser nos expérimentations, il a été nécessaire de : (1) rassembler un corpus déjà annoté en termes de difficulté ; (2) développer un extracteur d'UPs nominales.

3.1 Le corpus

La conception d'une formule de lisibilité implique l'annotation d'un corpus en terme de difficulté et, par conséquent, le choix d'une échelle de référence. Dans le contexte de l'enseignement des langues en Europe, le Cadre européen commun de références pour les langues (CECR) (Conseil de l'Europe, 2001) s'est imposé comme un choix évident. Cette norme définit six niveaux de progression – A1 (débutant) ; A2 ; B1 ; B2 ; C1 et C2 (avancé). Toutefois, afin de mieux rendre compte de l'évolution des apprenants, plus rapide dans les premières phases de l'apprentissage, nous avons scindé les trois premiers niveaux obtenant ainsi neufs échelons distincts.

Notons que, depuis son introduction, le CECR a amené une certaine standardisation des manuels de FLE. Il a dès lors été possible de rassembler, au départ d'un sous-ensemble de ces manuels, un corpus de textes annotés par des experts [2]. Nous avons ainsi rassemblé 1 895 textes, dont la taille varie de quelques phrases à plus de 2000 mots. Par ailleurs, afin de disposer d'un corpus de tests dont la probabilité *a priori* de chaque classe est similaire, nous avons sélectionné aléatoirement 50 textes par niveau, retenant ainsi 450 textes pour nos expérimentations.

3.2 L'extraction des UPs

En ce qui concerne la procédure d'extraction des UPs, elle s'inspire largement des travaux de Daille (1995) en ce qu'elle associe, à la validation statistique, un ensemble de filtres linguistiques. En plus d'assurer une plus grande précision, ces filtres nous permettent de contraindre la nature grammaticale des candidats termes et d'ainsi restreindre l'extraction aux structures nominales.

Dans ce système, la validation statistique implémente le rapport de la log-vraisemblance tel que décrit dans Silva & Lopes (1999). Le fonctionnement de cette mesure, comme de l'ensemble des mesures d'association, suppose une masse fréquentielle conséquente, masse qui n'était pas accessible au départ des textes de notre corpus. Pour palier ce problème, nous proposons dans Watrin & François (2011) le recours à une référence fréquentielle, c'est-à-dire une base de données de n-grammes (et de leur fréquence) construite au départ de gros corpus. Dans le cadre de cette étude, nous avons envisagé l'indexation des 5-grammes du corpus de Google (Michel *et al.*, 2011) (pour les années 2000-2008) et d'un corpus de 5 000 000 de mots issus de l'année 2009 du quotidien belge *Le Soir*.

4 Résultats et discussion

4.1 Capacité prédictive des unités polylexicales

Afin de déterminer l'intérêt des UPs en lisibilité du FLE, nous avons défini six variables à même de prendre en compte diverses facettes de ce phénomène : (1) la proportion d'unités polylexicales nominales par rapport au nombre de mots (**NCPW**) ; la proportion des 4 familles de structures morpho-syntaxiques suivantes : (2) *N N*, (3)

1. 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).

2. Bien entendu, ce processus de sélection suppose que le niveau d'un texte est similaire à celui du manuel dont il est extrait. Pour le détail des critères utilisés pour la sélection, consulter François (2009).

	Le Soir				Google			
Seuils θ	0	15	25	43	0	139	4000	9931
NCPW	$0,30^{3}$	$0,14^{2}$	$0,13^{2}$	$0,14^{2}$	$0,17^{3}$	$0,10^{1}$	$0,15^{2}$	$0,15^{2}$
NN	$-0,24^{3}$	$-0,14^{2}$	$-0,01$	$0,03$	$-0,22^{3}$	$-0,13^{2}$	$0,004$	$0,007$
NPN	$0,05$	$0,13^{2}$	$0,09$	$0,11^{1}$	$0,04$	$0,06$	$0,15^{2}$	$0,17^{3}$
AN	$-0,05$	$-0,03$	$0,02$	$0,08$	$-0,07$	$0,03$	$0,08$	$0,09^{1}$
NA	$0,36^{3}$	$0,30^{3}$	$0,27^{3}$	$0,22^{3}$	$0,37^{3}$	$0,32^{3}$	$0,25^{3}$	$0,28^{3}$
MeanP-UP	$-0,03$	$-0,03$	$-0,04$	$-0,05$	$0,15^{2}$	$0,16^{2}$	$0,14^{1}$	$0,09$

TABLE 1 – Corrélation entre les variables indépendantes et la difficulté des textes. Le taux de significativité est indiqué comme suit : $^{1}p < 0,05$; $^{2}p < 0,01$; $^{3}p < 0,0001$

N PREP (DET) N, (4) *A N* et (5) *N A* ; et (6) la probabilité moyenne d'apparition des unités polylexicales du texte (**MeanP-UP**), calculée sur la base des deux références décrites précédemment.

Par ailleurs, nous avons également fait varier le seuil θ associé à la validation statistique. De cette manière, nous avons pu estimer l'impact de la force d'association des constituants des UPs. Quatre seuils ont été retenus pour chacun des deux corpus de 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 (Watrin & François, 2011) et une valeur intermédiaire comme troisième seuil. La Table 1 rapporte les coefficients de corrélation de Pearson (r) entre les 6 variables susmentionnées et le niveau de difficulté des textes du corpus de tests.

Ces résultats apportent des enseignements intéressants. Premièrement, on note que plusieurs variables, en particulier **NPCW** et la structure NA, sont significativement associées avec la difficulté des textes de notre corpus. Elles semblent à première vue être de bons candidats pour prédire la difficulté de textes. Cependant, lorsqu'on les intègre au sein d'une formule de lisibilité de type classique, c'est-à-dire qui considère uniquement le nombre de lettres par mots (**NLM**) et le nombre de mots par phrase (**NMP**) [3], leur apport n'est pas significatif ($\chi^2 = 2,98$; $p-value = 0,08$) [4]. Les UPs n'amènent donc pas d'information véritablement nouvelle par rapport aux variables traditionnelles.

Une seconde observation notable est que le fait d'augmenter θ, et donc de renforcer le taux de cohésion des UPs, tend à diminuer l'efficacité des différents prédicteurs. Sur base de ces résultats, on pourrait en conclure que les UPs sont de moins bons prédicteurs que l'ensemble des structures nominales complexes ($\theta = 0$). Cependant, il nous semble plus juste de limiter ce constat d'échec aux UPs détectées automatiquement à l'aide de techniques statistiques. 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 ».

Confrontés à cette inadéquation des techniques de détection des UPs au contexte de la lisibilité, nous nous sommes posés une seconde question. Qu'en est-il de modèles plus simples, les modèles n-grammes, qui considèrent des suites de tokens sans motivation linguistique ?

4.2 Les modèles n-grammes

Pour vérifier l'efficacité de modèles n-grammes en lisibilité, nous nous sommes servis de la fréquence des n-grammes d'ordre 1 à 5 contenus dans nos références (*cf.* Section 3). Seul le modèle bigramme s'est avéré pertinent pour notre approche. En effet, la capacité de discrimination des modèles d'ordre supérieur souffre trop du lissage, le nombre de n-grammes inconnus augmentant proportionnellement à l'ordre du modèle. La probabilité des évènements inconnus étant toujours la même, les variables qui en découlent ne sont plus suffisamment discriminantes.

Sur la base des bigrammes de Google et du Soir, nous avons défini deux familles de prédicteurs. La première repose sur les probabilités conditionnelles des mots, $P(w_i|w_{i-1})$, desquels nous avons dérivé un modèle n-gramme classique, normalisé en fonction du nombre m de mots par texte selon la formule suivante :

3. Les deux formules de lisibilité utilisées pour cette comparaison reposent sur une régression logistique ordinale, qui est décrite en détail dans François (2009).

4. La technique statistique de comparaison entre les deux modèles assimile chacun d'eux à une hypothèse d'explication des données et en calcule le rapport de la log-vraisemblance multiplié par la constante -2. Dès lors, ce rapport est distribué selon une loi chi-carré.

	normTLProb	MeanProb	MedianProb	meanNGProb	medianNGProb	gmeanNGProb
Google	$0,003$	$0,33^3$	$-0,04$	$0,38^3$	$-0,001$	$-0,03$
Le Soir	$-0,06$	$0,18^2$	$0,01$	$0,25^3$	$-0,09$	$-0,0007$

TABLE 2 – Corrélation entre les variables basées sur les n-grammes et la difficulté des textes.

$$\text{normTLProb} = \frac{1}{m}\sum_{i=1}^{m} \log P(w_i|w_{i-1})$$

En marge de ce modèle, nous avons aussi considéré la moyenne des probabilités conditionnelles (**MeanProb**) et leur médiane (**MedianProb**). La seconde famille de prédicteurs s'intéresse quant à elle directement aux probabilités des bigrammes, $P(w1 \cap w2)$, qui correspondent davantage aux probabilités prises en compte dans notre approche des UPs. Nous avons ainsi calculé la moyenne (**meanNGProb**), la médiane (**medianNGProb**) et la moyenne géométrique (**gmeanNGProb**), qui comme un modèle n-gramme classique, multiplie les probabilités entre-elles plutôt que d'en faire la somme. Les corrélations de ces 6 variables avec la difficulté sont reprises dans la Table 2.

Là aussi, nos analyses apportent quelques constatations d'intérêt. La première d'entre-elles est l'inefficacité complète des variables basées sur un modèle bigramme classique (r vaut $0,003$ et $-0,06$ pour **normTLProb**), ce qui nous paraît hautement surprenant en regard des résultats rapportés dans les travaux sur l'anglais. Schwarm & Ostendorf (2005), par exemple, disent employer des modèles n-grammes avec succès. Notons qu'ils ne rapportent pas de corrélation individuelle pour cette variable et que leurs bonnes performances globales sont obtenues à l'aide de nombreux prédicteurs.

Par contre, la moyenne des probabilités des bigrammes du texte (**MeanProb**) apparaît largement significative. D'ailleurs, son addition à la formule de lisibilité baseline envisagée précédemment amène cette fois une amélioration significative des performances ($R = 0,67$; $\chi^2 = 11,66$; $p - value = 0,0006$). Il est particulièrement intéressant par rapport à notre démarche de noter que **MeanProb** est plus informatif qu'une variable plus fine, qui calcule la moyenne des probabilités des UPs (**MeanP-UP**). Ce résultat nous semble mettre sérieusement en question l'intérêt des UPs en lisibilité, et vient par ailleurs répliquer les résultats de Ozasa *et al.* (2007) sur l'anglais.

5 Conclusion

Dans cette étude, nous avons testé les apports de la notion d'UPs dans une formule de lisibilité pour le FLE. Ceux-ci se sont révélés négligeables, aussi bien de façon absolue qu'en comparaison avec une approche plus simple basée sur les n-grammes. Notre expérience souligne que la prise en compte automatique de notions plus linguistiques ne conduit pas à des résultats satisfaisants pour la lisibilité du FLE. En effet, les traitements TAL nécessaires sous-tendent trop d'approximations qui nuisent à l'efficacité des variables : erreurs de l'extracteur d'UPs, problème de couverture des références, etc.

Par ailleurs, nos expérimentations ont jeté un doute sur l'efficacité des modèles n-grammes classiques pourtant largement employés dans le domaine. En effet, rappelons que, sur notre corpus, une simple approche par unigramme produit une corrélation élevée ($r = -0,59$), ce qui n'est pas le cas des modèles d'ordre supérieur (bigramme : $r = -0,06$). Il nous faut donc conclure soit à un problème d'estimation des probabilités au niveau des références, soit remettre en question 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 requièrerait cependant des expérimentations dans le cadre de modèles explicatifs et non prédictifs.

Enfin, on peut se demander si ce motif de résultats se reproduirait si (1) les UPs verbales étaient également considérées ; si (2) le repérage des UPs était effectué manuellement (ce qui représente toutefois un travail énorme) ; et si (3) seules les expressions figées, plus opaques sémantiquement, étaient prises en compte. Cette dernière perspective, intellectuellement attrayante, doit être relativisée, car il est fort probable que ce type d'UPs soit trop rare dans les textes à analyser pour se révéler statistiquement informatif.

Références

CONSEIL DE L'EUROPE (2001). *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*. Paris : Hatier.

DAILLE B. (1995). *Combined approach for terminology extraction : lexical statistics and linguistic filtering*. Rapport interne, Lancaster University.

DIAS G., GUILLORÉ S. & 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*, p. 213–221.

FRANÇOIS T. (2009). 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*, p. 19–27.

HOWELL D. (2008). *Méthodes statistiques en sciences humaines, 6ème édition*. Bruxelles : De Boeck.

KODA K. (2005). *Insights into second language reading : A cross-linguistic approach*. Cambridge : Cambridge University Press.

LAROCHE J. (1979). Readability measurement for foreign-language materials. *System*, **7**(2), 131–135.

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. & AIDEN E. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, **331**(6014), 176–182.

OZASA T., WEIR G. & FUKUI M. (2007). Measuring readability for Japanese learners of English. In *Proceedings of the 12th Conference of Pan-Pacific Association of Applied Linguistics*.

PAWLEY A. & SYDER F. (1983). Two puzzles for linguistic theory : nativelike selection and nativelike fluency. In J. RICHARDS & R. SCHMITT, Eds., *Language and Communication*, p. 191–225. London : Longman.

SCHWARM S. & 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*, p. 523–530.

SILVA J. & 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*.

UITDENBOGERD S. (2005). Readability of French as a foreign language and its uses. In *Proceedings of the Australian Document Computing Symposium*, p. 19–25.

UNDERWOOD G., SCHMITT N. & GALPIN A. (2004). The eyes have it : An eye-movement study into the processing of formulaic sequences. In N. SCHMITT, Ed., *Formulaic sequences : acquisition processing and use*, p. 155–172. Amsterdam : John Benjamins.

WATRIN P. & FRANÇOIS T. (2011). N-gram frequency database reference to handle MWE extraction in NLP applications. In *Unpublished manuscript*.

WEIR G. & ANAGNOSTOU N. (2008). Collocation frequency as a readability factor. In *Proceedings of the 13th Conference of the Pan Pacific Association of Applied Linguistics*.