Les apports du TAL à la lisibilité du français langue étrangère

Thomas François — Cédrick Fairon

Centre de Traitement automatique du langage, ILC (UCLouvain)
Place Blaise Pascal, 1 bte L3.03.12
1348, Belgium
thomas.francois@uclouvain.be, cedrick.fairon@uclouvain.be

RÉSUMÉ. *Cet article décrit une série d'expériences visant à (1) évaluer la contribution des techniques TAL à la mesure de la lisibilité des textes du français langue étrangère (FLE) et (2) à proposer une nouvelle formule de lisibilité spécifique au FLE. Ce nouveau modèle utilise quarante-six variables qui modélisent diverses caractéristiques lexicales, syntaxiques et sémantiques des textes, ainsi que certaines particularités du contexte FLE. L'article présente également une série de comparaisons entre des techniques de sélection de variables et des algorithmes d'apprentissage automatisé. Il apparaît que notre meilleur modèle, fondé sur les machines à vecteurs de support (SVM), surpasse de manière significative les modèles précédents pour le français. Quant à la contribution des techniques TAL à la lisibilité, nos résultats suggèrent que l'usage de variables TAL au sein des modèles ne produit pas des résultats significativement supérieurs à une approche classique, mais que combiner les deux types d'information conduit à une amélioration significative des performances.*

ABSTRACT. *This paper presents a set of experiments aiming to (1) assess the contribution of NLP to the specific issue of the readability of texts for French as a foreign language (FFL) readers and (2) to propose a new readability formula for FFL. This new model relies on 46 textual features representative of the lexical, syntactic, and semantic levels as well as some of the specificities of the FFL context. We report comparisons between several techniques for feature selection and for various learning algorithms. Our best model, based on support vector machines (SVM), significantly outperforms previous FFL formulas. Regarding the contribution of NLP to readability, our findings suggest that NLP-based models are not significantly better than the classic ones, although combining both type of information leads to significant improvment.*

MOTS-CLÉS : *lisibilité du FLE, traitement automatique du langage, difficulté des textes.*

KEYWORDS: *readability of FFL, natural language processing, difficulty of texts.*

TAL. Volume 54 – n° 1/2013, pages 171 à 202

1. Introduction

Depuis des millénaires, la lecture constitue une voie privilégiée pour entrer en contact avec d'autres cultures et d'autres époques, permettant ainsi de faire l'expérience de l'altérité. Ces dernières années, le développement de l'Internet a rendu particulièrement aisé l'accès à un grand nombre d'écrits, favorisant par la même occasion les contacts entre langues et l'exposition du public à cette diversité linguistique.

Dès lors, il n'est pas étonnant que l'enseignement des langues étrangères et secondes (L2) soit une préoccupation particulièrement importante dans nos sociétés. L'Union Européenne a notamment traduit cette préoccupation *via* la publication d'un Cadre européen commun de référence pour les langues (CECR) (Conseil de l'Europe, 2001), qui vise à baliser l'enseignement des L2 au sein de l'Europe. Ce référentiel défend une perspective pédagogique appelée « actionnelle », qui accorde au texte et à la lecture un rôle essentiel dans l'acquisition des langues. En effet, diverses expérimentations ont montré que la pratique régulière de la lecture permet non seulement d'améliorer ses compétences à ce niveau (Reitsma, 1988), mais favoriserait aussi plus globalement l'acquisition de la langue étrangère (Krashen, 1989).

Pour que cette pratique soit profitable, encore faut-il que les textes soient adaptés au niveau des étudiants (O'Connor *et al.*, 2002), ce qui n'est pas toujours le cas, d'après Mesnager (1986), qui prend l'exemple de la littérature de jeunesse francophone. Le problème est que trouver un texte sur un sujet précis et adapté au niveau d'une classe requiert du temps, ce dont les professeurs de langue ne disposent pas toujours. C'est là qu'intervient la lisibilité, un domaine qui, depuis les années 1920, vise à mettre au point des techniques reproductibles, capables d'associer un texte à un lecteur sans que soit nécessaire l'intervention d'un jugement humain à propos du niveau de difficulté dudit texte.

Ces techniques ont d'abord pris la forme d'une équation mathématique relativement simple (voir section 2) dont les variables indépendantes étaient fondées sur des comptages, de mots, de syllabes ou de lettres. On y réfère souvent *via* l'appellation « formules classiques ». Plus récemment, diverses techniques issues du TAL et de l'apprentissage automatisé sont venues modifier et renforcer l'approche traditionnelle. Elles recourent notamment à des modèles de langue (Schwarm et Ostendorf, 2005 ; Collins-Thompson et Callan, 2005), des analyseurs syntaxiques (Schwarm et Ostendorf, 2005), diverses mesures de la cohérence et de la structure des textes (Feng *et al.*, 2010 ; Pitler et Nenkova, 2008) ou l'analyse des unités polylexicales (François et Watrin, 2011). La majorité de ces études portent sur l'anglais langue première (L1) et estiment que l'application de techniques TAL en lisibilité permet d'obtenir de meilleurs résultats que l'approche classique.

Dans cet article, nous articulons une série d'expériences autour de deux objectifs principaux : (1) vérifier si, dans un contexte donné, à savoir la lisibilité du français langue étrangère (FLE), cet apport du TAL et de l'apprentissage automatisé se confirme bien ; (2) proposer une nouvelle formule de lisibilité pour le FLE, un do-

maine très peu exploré du point de vue de la lisibilité et qui n'a notamment pas encore bénéficié des apports des techniques de TAL et d'apprentissage automatisé.

Dans la section 2, nous détaillons notre problématique, en explorant brièvement les différents courants en lisibilité, avant de faire le tour des travaux spécifiques à la lisibilité du FLE. Ensuite, la section 3 expose les étapes méthodologiques relatives à la conception de cette nouvelle formule de lisibilité, dont les résultats sont rapportés dans la section 4. La section 5 revient plus en détail sur la question des apports du TAL à la lisibilité, présentant des expérimentations supplémentaires à ce niveau.

2. La lisibilité et le TAL

2.1. *La lisibilité avant le TAL*

La première formule de lisibilité remonte aux travaux de Lively et Pressey (1923), même si c'est véritablement à Vogel et Washburne (1928) que l'on doit le développement de la méthodologie utilisée tout au long du XX^{e} siècle. Elle modélise la difficulté sous la forme d'une équation de régression linéaire multiple, dans laquelle la difficulté du texte est la variable dépendante à prédire. Les variables indépendantes sont autant de facteurs textuels que le chercheur estime être de bons prédicteurs de la difficulté. Pour exemple, voici l'une des formules les plus populaires, due à Flesch (1948, 225) :

$$\text{Reading Ease} = 206{,}835 - 0{,}846\, wl - 1{,}015\, sl \qquad [1]$$

où :

Reading Ease (RE) : un score compris entre 0 (très difficile) et 100 (équivaut à un texte au sujet duquel un enfant de 4^{e} primaire [1] obtiendrait un résultat de 75 % à un test de compréhension) ;

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

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

Cette formule a été obtenue *via* une régression linéaire multiple appliquée à un corpus de textes issus des *Standard Test Lessons in Reading* de McCall et Crabbs. Elle est représentative de l'approche dite « classique » en lisibilité, qui a donné naissance aux célèbres formules de Flesch (1948), Dale et Chall (1948) ou encore Fry (1968). Ce paradigme s'appuie sur les observations de Lorge (1944) selon qui les différents facteurs de la difficulté sont fortement intercorrélés. En conséquence, au vu du temps nécessaire pour le comptage des variables – lequel s'effectuait manuellement à l'époque –, les formules classiques visent à obtenir des modèles efficaces avec très peu de variables, généralement deux.

Dans les années 70, le développement de la psychologie cognitive conduit de nombreux chercheurs à s'opposer virulemment à la lisibilité traditionnelle (Kintsch et

1. Cela équivaut à la classe de CM1 en France.

Vipond, 1979 ; Redish et Selzer, 1985). Ils estiment que les formules classiques se fondent uniquement sur des caractéristiques de surface (ex. le nombre de mots, de lettres, etc.) et omettent des dimensions de plus haut niveau qui expliqueraient, au moins autant, la difficulté d'un texte. De plus, ces formules font aussi abstraction de l'aspect interactif du processus de lecture. Dès lors, dans la lignée de ces critiques, apparaît un nouveau paradigme, dit structuro-cognitiviste. Il vise à dépasser les limitations précitées de l'approche classique et se concentre sur les dimensions structurelles et cognitives des textes, telles que la charge inférentielle (Kintsch et Vipond, 1979 ; Kemper, 1983), la densité conceptuelle (Kintsch et Vipond, 1979), ou la macrostructure (Meyer, 1982).

Cependant, malgré cet usage de variables plus sophistiquées – et par conséquent plus complexes à mettre en œuvre –, les modèles structuro-cognitivistes ne semblent pas capables de dépasser les résultats obtenus par l'approche traditionnelle. Ainsi, Kemper (1983) compare sa propre formule, qui modélise la charge inférentielle d'un texte à l'aide de trois variables, à celle de Dale et Chall (1948), qui repose sur le nombre de mots absents d'une liste de mots simples et sur le nombre moyen de mots par phrase. Kemper en conclu que les deux approches sont similaires en termes de pouvoir prédictif (R, le coefficient de corrélation multiple des variables avec la difficulté des textes est de $0{,}76$). De plus, lorsqu'elle combine ses variables cognitives avec les deux variables classiques de Dale et Chall (1948), la formule obtenue n'est guère plus efficace ($R = 0{,}78$).

2.2. *Le dilemme des années 90 et l'arrivée du TAL en lisibilité*

Dans les années 90, les chercheurs se trouvent confrontés à un dilemme : continuer à expérimenter les aspects structuro-cognitivistes – souvent difficiles à paramétriser et parfois impossibles à implémenter – ou revenir à une approche traditionnelle et prêter ainsi le flanc à la critique. Ce n'est dès lors pas une coïncidence si la quantité de recherches en lisibilité diminue dans les années 90.

À la fin des années 90, on voit apparaître les premières tentatives pour résoudre ce dilemme grâce à l'intervention de techniques de TAL. Ainsi, Foltz *et al.* (1998) appliquent l'analyse sémantique latente pour estimer la cohérence d'un texte, montrant qu'il est possible de prendre en compte des variables complexes de façon complètement automatisée. Peu après, Si et Callan (2001) suggèrent une transformation radicale de la méthodologie classique en reformulant la question de la lisibilité des textes sous la forme d'un problème de classification – plutôt qu'une régression – et en lui appliquant des techniques issues de l'apprentissage automatisé, en l'occurrence un modèle bayésien naïf. Ils font aussi usage d'un modèle de langue unigramme qui modélise la distribution du lexique par niveau de difficulté.

2.3. *TAL et lisibilité*

Ces travaux relancent le domaine, qui fait dès lors davantage l'objet de recherches en TAL qu'en psychologie de l'éducation. Ce courant, que François (2011a) regroupe sous le terme de « lisibilité computationnelle », est caractérisé par trois traits principaux : (1) le recours à des algorithmes d'apprentissage automatisé – généralement de classification – pour remplacer la régression linéaire, (2) une sophistication des variables linguistiques prises en compte dans les textes grâce à l'emploi de techniques de TAL et (3) le recours à un large corpus de textes.

Parmi les principaux travaux de ce paradigme, citons Collins-Thompson et Callan (2005), qui perfectionnent le classifieur bayésien naïf de Si et Callan (2001). Ce faisant, ils parviennent à de meilleurs résultats que les formules traditionnelles, du moins sur des textes issus du Web. Simultanément, Schwarm et Ostendorf (2005) développent un classifieur par machine à vecteurs de support (SVM), qui combine un modèle trigramme et diverses variables dérivées d'un analyseur syntaxique. Là aussi, leur modèle se compare favorablement par rapport à une formule classique, celle de Kincaid *et al.* (1975). Ces investigations se limitent aux dimensions lexicales et syntaxiques, mais suggèrent qu'à ce niveau, l'emploi de variables TAL et de modèles d'apprentissage statistique améliore la qualité des prédictions.

Il faut attendre Pitler et Nenkova (2008) pour voir les aspects discursifs être explorés plus en détail selon une approche TAL. Les auteurs recourent au *Penn Discourse TreeBank* (Prasad *et al.*, 2008), où les types de relation entre les phrases et les connecteurs sont annotés, et en dérivent plusieurs variables, telles que la vraisemblance des relations discursives au sein d'un texte. Dans la même lignée, Feng *et al.* (2009) explorent diverses variables cognitives dans le cadre de lecteurs adultes atteints de déficiences intellectuelles. En effet, ce sont souvent les limitations mémorielles qui sont la cause des difficultés rencontrées par ce type de public. Les auteurs définissent un ensemble de variables fondées sur les « entités » (qui rassemblent les noms propres et les noms communs) et sur la notion de chaînes lexicales, décrite par Galley et McKeown (2003). Ils montrent que c'est le modèle combinant des variables classiques et les variables cognitives qui se comporte le mieux sur des textes de presse. Par ailleurs, les aspects cognitifs se révèlent plus efficaces dans le cas des adultes atteints de déficiences intellectuelles. Toutefois, sur une tâche plus générale, Feng *et al.* (2010) concluent à la relative inefficacité des variables de discours. Le bilan est donc mitigé pour cette classe de variables, que les structuro-cognitivistes avaient pourtant mis sur un piédestal.

Si l'approche TAL apparaît profitable dans ces études, certains éléments viennent jeter le doute sur cette conclusion. Ainsi, Collins-Thompson et Callan (2005) rapportent que les variables classiques *type token ratio* (TTR) et « nombre de mots absents de la liste de 3 000 mots de Dale » obtiennent de meilleures corrélations que leur modèle de langue sur un corpus de livres simplifiés pour enfants. Sur d'autres langues que l'anglais, Aluisio *et al.* (2010) obtiennent comme meilleur prédicteur le nombre moyen de mots par phrase. Ces résultats sont assez étonnants et ils soulèvent

la question de l'intérêt d'employer des variables TAL complexes, sachant que celles-ci mobilisent une quantité de ressources computationnelles largement supérieure. C'est pourquoi, nous avons décidé d'explorer cette question dans un contexte qui n'avait que très peu été exploré précédemment : la lisibilité du FLE.

2.4. *La lisibilité du FLE*

Les formules de lisibilité, d'abord développées pour l'anglais L1 dans un contexte scolaire, ont peu à peu été adaptées à d'autres langues et à d'autres contextes, parmi lesquels la lecture en langue étrangère. Or, lorsqu'on transpose une formule d'une population à une autre, il convient, *a minima*, d'adapter les coefficients des variables employées, voire de changer les variables prises en compte. Kandel et Moles (1958), les auteurs de la première formule pour le français L1, observent ainsi qu'il convient de modifier les coefficients de la formule de Flesch avant de l'appliquer au français, puisque les mots sont en moyenne 1,15 fois plus longs dans les textes francophones que dans les textes anglophones.

En ce qui concerne le FLE, il apparaît d'autant plus nécessaire d'adapter les formules que les processus de lecture en L1 et en L2 diffèrent sensiblement. Koda (2005) souligne ainsi que le lecteur en L2 ne peut se reposer sur une connaissance orale de la langue lorsqu'il apprend à lire. Il est donc plus sensible à des difficultés au niveau du lexique et de la syntaxe. En revanche, il subit l'influence d'une ou plusieurs langues préexistantes. Tharp (1939) défendait déjà l'importance de cette adaptation et suggérait d'adapter les mesures de la complexité lexicale en prenant en compte l'effet facilitateur des congénères [2].

Les travaux sur le FLE ne se développent toutefois qu'à partir des travaux de Cornaire (1988), qui étudie les possibilités d'adaptation des formules de Henry (1975) – développées pour le français langue maternelle – au FLE. Étant donné que Henry a développé trois formules, respectivement pour les élèves de la 5e et 6e primaires, de la 2e et 3e secondaires et de la 5e et 6e secondaires [3], Cornaire en vient à déterminer une équivalence entre ce système scolaire belge et les niveaux utilisés pour le FLE au Québec à l'aide de tests de closure [4]. L'auteur conclut à la possibilité d'employer cette formule pour le FLE, avec certaines réserves.

2. Il s'agit de mots d'une L2 dont la forme est homographique ou quasi homographique avec leurs équivalents sémantiques dans la L1 du lecteur.

3. Dans le système français, ces groupes correspondent respectivement à la classe de CM2 et la 6e , à la 4e et la 3e et à la 1re et la terminale.

4. Le test de closure a été développé par Taylor (1953). Il s'agit d'un texte à trous où les blancs sont réguliers et qui est supposé mesurer le niveau de compréhension du texte par un lecteur.

Plus récemment, Uitdenbogerd (2005) revient à la question des congénères et propose une formule (*FR*) qui combine le nombre moyen de mots par phrase (*WpS*) et le nombre de congénères par 100 mots (*Cog*) :

$$FR = 10 * WpS - Cog \quad [2]$$

Cette formule se démarque de celle de Tharp (1939) par l'ajout d'un facteur syntaxique, mais reste classique dans sa méthodologie. De plus, elle se limite à un public d'apprenants anglophones du FLE, car elle a non seulement été entraînée sur des données spécifiques à ce public, mais elle nécessite également de détecter les congénères spécifiques à la paire de langues français-anglais.

C'est pourquoi François (2009b) a récemment publié une « formule computationnelle » pour le FLE, laquelle vise un public plus large, car elle ne postule pas d'hypothèse sur la langue maternelle des lecteurs. Son modèle, qui recourt à la régression logistique et à une dizaine de variables, obtient des résultats intéressants, mais qui nécessitent d'être améliorés pour permettre un emploi réel du modèle en FLE. Par ailleurs, l'étude n'explore qu'un ensemble réduit de variables (une vingtaine). Dans la suite de cet article, nous proposons une étude de plus grande ampleur, visant elle aussi à développer une formule computationnelle pour le FLE, sur la base d'un large ensemble de variables linguistiques.

3. Méthodologie de conception de la formule

Dans la majorité des cas, la conception d'une formule de lisibilité computationnelle s'apparente à un problème de classification et comporte trois étapes. Tout d'abord, il s'agit de rassembler un corpus dont la difficulté des textes est connue. La méthode de récolte des données que nous avons employée est décrite à la section 3.1. L'étape suivante, détaillée à la section 3.2, consiste à définir un ensemble de prédicteurs, c'est-à-dire des variables linguistiques utiles pour prédire la difficulté des textes du corpus. Enfin, il convient d'entraîner un modèle de classification sur le corpus, ce qui demande également d'identifier le meilleur sous-ensemble de prédicteurs. Pour des raisons de clarté de l'exposé, cette dernière étape a été divisée entre la présentation des algorithmes statistiques utilisés (section 3.3) et la description des expérimentations concernant la sélection des variables (section 4).

3.1. *Le corpus de référence*

En lisibilité, un corpus de référence est un ensemble de textes destinés à la population ciblée par la formule et dont le niveau de difficulté pour cette population est connu et exprimé selon une échelle de difficulté donnée. Traditionnellement, pour obtenir une telle annotation, on recourt à des tests de compréhension ou à des tests de closure. Cependant, dans le cadre de la lisibilité computationnelle, le recours à des techniques de TAL et à des algorithmes d'apprentissage plus complexes rend nécessaire de disposer d'un grand nombre de textes. Par conséquent, il n'est pas possible

de tester directement la compréhension d'un échantillon de lecteurs issus de la population d'intérêt. On recourt alors à des textes dont la difficulté a été jugée par des experts, typiquement disponibles dans les manuels scolaires ou les livres simplifiés.

Dans notre cas, s'ajoutait à ce besoin, le désir d'utiliser une échelle de difficulté pratique pour les utilisateurs de notre formule. Notre choix s'est dès lors tout naturellement porté vers l'échelle du Cadre européen commun de référence pour les langues, qui comporte six niveaux : A1 (niveau découverte), A2 (niveau de survie), B1 (niveau seuil), B2 (niveau indépendant), C1 (niveau autonome) et C2 (niveau maîtrise). Cette échelle est en effet devenue la référence dans le milieu de l'enseignement des langues étrangères en Europe, dans lequel évoluent les utilisateurs potentiels de notre formule.

Cette échelle présente un autre avantage. Depuis son introduction en 2001, les nouveaux manuels de FLE se positionnent en référence à cette échelle et il est dès lors possible de les utiliser comme source de textes annotés. Nous avons ainsi retenu vingt-huit manuels de FLE sur la base des critères suivants :

– les manuels doivent utiliser l'échelle du CECR et donc être postérieurs à 2001. Cela permet aussi de s'assurer de la modernité de la langue modélisée ;

– seuls des manuels destinés à des adultes ou à des jeunes sont retenus, puisqu'il s'agit de la population de lecteurs visée par notre formule ;

– notre formule vise à modéliser la lecture enseignée dans une classe dispensant un enseignement généraliste du FLE, nous avons donc rejeté les manuels de français sur objectifs spécifiques (FOS).

Notons aussi qu'au sein des manuels retenus, seuls les textes associés à une tâche de compréhension à la lecture ont été sélectionnés, ce qui nous a permis de rassembler près de 2 160 textes. Représentant plus de 500 000 mots, ceux-ci couvrent des sujets divers : extraits de littérature, articles de journaux, dialogues, recettes de cuisine, etc. Il importe, enfin, de souligner que chaque texte s'est vu attribuer le même niveau de difficulté que le manuel dont il est issu.

Si cette approche par « jugements d'experts » constitue aujourd'hui la norme en lisibilité computationnelle, elle ne va pas sans sérieuses limitations. En effet, van Oosten *et al.* (2011) ont mis en évidence que les jugements d'experts étaient sujets à une grande variance et que l'accord interjuges pouvait se révéler insuffisant pour une tâche de classification. Dans notre cas, les niveaux de chaque manuel sont définis par leurs auteurs, lesquels changent d'une collection à l'autre, voire au sein d'une même collection. Par conséquent, en accord avec les résultats de van Oosten *et al.* (2011), nous avons noté des différences de jugements substantielles entre les différentes séries de manuels de notre corpus. Afin d'évaluer plus précisément cette hétérogénéité, nous avons calculé, pour chacun des vingt-huit manuels, son nombre moyen de lettres par mot et son nombre moyen de mots par phrase. Bien que ces deux variables ne représentent qu'imparfaitement la difficulté des textes, leur efficacité a été prouvée à maintes reprises dans la littérature. À partir de ces données, nous avons effectué un test d'analyse de la variance (ANOVA) pour chacun des niveaux, afin de

	A1	A2	B1	B2	C1	C2	Total
Nb. de textes	430	380	552	198	184	108	1 852
Nb. de mots	58 561	75 779	176 973	71 701	92 327	35 202	510 543

Tableau 1. *Distribution du nombre de textes et de mots par niveau dans notre corpus*

déterminer la cohérence des annotations à l'intérieur d'un niveau. Cette hétérogénéité a été détectée dans trois des six niveaux (A1, A2 et B1) et une analyse qualitative subséquente a révélé qu'une proportion importante de cette hétérogénéité provenait de manuels (ex. *Rond-Point*) appliquant la nouvelle approche didactique recommandée par le CECR : l'approche par les tâches. En effet, ce type de manuels accorde plus d'importance à la tâche associée au texte qu'à celui-ci lorsqu'ils attribuent un niveau de difficulté à la combinaison des deux. Par conséquent, nous avons résolu d'écarter de notre corpus les cinq manuels (pour un total de 249 textes) qui ressortaient de ce paradigme. Les 1 852 documents restants furent utilisés pour nos expérimentations. Leur distribution par niveau est reprise au tableau 1. Signalons que cette solution a permis d'améliorer l'homogénéité du corpus, sans régler complètement le problème.

3.2. *Les prédicteurs*

Une fois le corpus de référence rassemblé, l'étape suivante consiste à identifier un ensemble de caractéristiques linguistiques, aussi appelées prédicteurs. Celles-ci doivent entretenir une relation de corrélation significative, voire une relation causale, avec la difficulté des textes. Dans le premier cas, on parle d'indices de la difficulté ; dans le second, de causes. Pour certains chercheurs, seuls les facteurs causals devraient être utilisés, malgré que les indices aient largement fait la preuve de leur efficacité. Dans notre étude, nous avons toutefois envisagé ces deux types de prédicteurs, en nous reposant sur deux sources d'information. D'une part, nous avons répliqué le plus possible de variables issues de la littérature en lisibilité de l'anglais et du français. D'autre part, nous avons exploré la littérature décrivant le processus de lecture en L1 et en L2 afin d'y découvrir un certain nombre de nouvelles idées de variables. Notre hypothèse est qu'un élément du texte ayant la propriété d'inhiber ou de faciliter la lecture est une variable potentiellement intéressante (dans la mesure où il est paramétrisable). De ces deux sources, nous avons tiré 406 variables.

Toutefois, un bon prédicteur doit également répondre à d'autres conditions, et, en particulier, être le moins corrélé possible avec les autres prédicteurs afin d'éviter des redondances d'informations. Dans cette optique, nous avons classé nos 406 variables en quatre familles, qui correspondent au type d'information qu'elles sont supposées apporter au modèle de lisibilité : lexical, syntaxique, sémantique et spécifique au FLE. Chacune d'elles a ensuite été subdivisée en un certain nombre de sous-familles, décrites en détail dans le reste de cette section. Nous indiquons également, à côté des noms de chaque famille, si celle-ci se rapproche plutôt des variables classiques (^{C})

ou des variables fondées sur le TAL (^{T}), une information utilisée dans les expériences rapportées à la section 5 [5].

3.2.1. *Les variables lexicales*

De nombreuses études ont montré que le niveau lexical constituait la source d'information la plus importante en lisibilité (Chall et Dale, 1995). Aujourd'hui, on sait que ces facteurs agissent essentiellement au niveau de la reconnaissance des mots, c'est-à-dire l'étape par laquelle un ensemble de traits sont reconnus comme un mot appartenant au lexique mental d'un individu (Lupker, 2005). Parmi ces facteurs, nous avons considéré la fréquence lexicale, la vraisemblance textuelle, la familiarité, la diversité lexicale, le voisinage orthographique et la longueur des mots.

3.2.1.1. Les statistiques de la fréquence lexicale$^{(C)}$

Les variables fondées sur la fréquence des mots reposent sur l'hypothèse d'une forte association entre la fréquence des mots dans la langue (ou fréquence objective) et leur facilité de lecture : l'effet de fréquence (Howes et Solomon, 1951 ; Monsell, 1991). Pour des raisons pratiques, celui-ci a longtemps été abordé *via* diverses simplifications. La technique la plus courante, employée entre autres par Dale et Chall (1948), consiste à définir une liste de mots simples et à calculer le pourcentage de mots absents de cette liste pour un texte donné. Les approches qui recourent directement à des fréquences calculées sur un large corpus sont plus récentes (Stenner, 1996 ; Graesser *et al.*, 2004).

Nous avons décidé de répliquer ces deux tendances : la première étant plus populaire, la seconde théoriquement plus précise. En ce qui concerne le pourcentage d'absents, deux listes ont été employées : la liste de Gougenheim *et al.* (1964), dans sa version longue [6], et une liste de 1 803 mots tirés des niveaux A1 et A2 du manuel FLE *Alter Ego* (Berthet *et al.*, 2006a ; Berthet *et al.*, 2006b) Dale et Chall (1948) ayant insisté sur la variation du pouvoir discriminant des variables reposant sur une liste en fonction de la taille de cette liste, différentes tailles de listes ont été testées, en particulier pour la liste de Gougenheim. Enfin, le comptage s'est fait tantôt au niveau de la forme en elle-même (*type)*, ou de ses instances (*tokens*), ce qui donne au total vingt-six variables différentes.

En ce qui concerne les statistiques lexicales, elles ont été obtenues à partir de la base fréquentielle Lexique 3 (New *et al.*, 2007), qui comprend environ 50 000 lemmes et 125 000 formes fléchies dont les fréquences ont été calculées sur un corpus de sous-

5. Il est évident que cette distinction est partiellement subjective et est critiquable. Le critère principal que nous avons utilisé est la présence de cette variable dans des études dites classiques en lisibilité. Cela explique pourquoi les variables reposant sur des ratios de catégories grammaticales sont considérées comme classiques, tandis que les variables qui utilisent les formes verbales sont vues comme venant du TAL, alors que toutes les deux nécessitent le recours à un taggeur.

6. Distribuée sur le site de Lexique 3 : www.lexique.org/public/gougenheim.php

titres de films de plus de 50 millions de mots. Ces fréquences ont été lissées *via* l'algorithme de Simple Good-Turing (Gale et Sampson, 1995) afin d'attribuer une valeur aux mots hors vocabulaire. Nous avons expérimenté diverses statistiques (moyenne, médiane, 75^{e} percentile, 90^{e} percentile, etc.) sur différentes formes (formes fléchies, lemmes), ainsi que sur diverses catégories grammaticales (seulement les noms, seulement les verbes, etc.). Au final, nous obtenons 84 variables.

3.2.1.2. La vraisemblance textuelle$^{(T)}$

Cette famille de variables repose également sur l'effet de fréquence, mais l'aborde à l'aide d'un modèle typiquement TAL : les n-grammes. Ceux-ci visent à estimer la probabilité qu'une séquence de mots existe dans la langue. La « lisibilité computationnelle » a utilisé les modèles n-grammes de deux façons différentes. Les premières investigations (Si et Callan, 2001 ; Collins-Thompson et Callan, 2005) emploient un modèle unigramme pour chaque niveau de difficulté et classent les textes en fonction de leur vraisemblance étant donné ces modèles. Plus récemment, d'autres ont employé comme variable un seul modèle, tantôt unigramme (Pitler et Nenkova, 2008 ; François, 2009b), tantôt trigramme (Kate *et al.*, 2010).

Dans cette étude, nous avons considéré un modèle unigramme, toujours fondé sur les probabilités lissées de Lexique 3 (lemme et formes fléchies), ainsi qu'un modèle bigramme, qui repose sur deux références : la base de n-grammes de Google (Michel *et al.*, 2011) et un corpus d'articles tirés du journal *Le Soir* constitué d'environ $5\ 000\ 000$ mots [7]. Dans les deux cas, la probabilité d'un texte donné $P(T)$ est normalisée en fonction du nombre n de mots :

$$P(T) = \frac{1}{n} \sum_{i=1}^{n} \log P(w_i|h) \qquad [3]$$

où w_i est le n-$i^{\text{ème}}$ mot et h, un historique limité de longueur 0 (unigramme) ou 1 (bigramme). Pour les bigrammes, nous avons également considéré des statistiques fondées sur les probabilités composées $P(w_i \cap h)$, telles que la moyenne, la médiane ou les 75^{e} et 90^{e} percentiles [8]. Au total, nous obtenons vingt et une variables.

3.2.1.3. La diversité lexicale$^{(C)}$

Un autre effet mis en évidence lors de l'étape de la reconnaissance des mots est l'effet de répétition des mots (Bowers, 2000). Il prédit que la reconnaissance d'un mot est plus aisée si ce dernier a été rencontré précédemment dans le texte. Par extension, moins le vocabulaire d'un texte est varié, plus celui-ci est supposé facile. C'est ce

7. Dans le cas du bigramme, le lissage a été effectué par interpolation linéaire (Chen et Goodman, 1999, 312).

8. Soulignons que des expérimentations ont également été conduites avec un modèle trigramme. Cependant, la proportion très élevée d'événements inconnus rend alors la majorité de ces variables égales – à cause du lissage – et par conséquent, non discriminantes.

phénomène que vise à capturer le classique TTR (*type token ratio*), utilisé dès Lively et Pressey (1923) en lisibilité. Celui-ci souffre cependant d'une limitation importante : sa normalisation. Dans cette étude, nous avons dès lors comparé le TTR classique, fondé sur les lemmes ou les formes fléchies, avec une version normalisée, qui calcule un TTR moyen pour chaque section de n mots (où n a été expérimentalement fixé à 100). Quatre variables supplémentaires ont ainsi été définies.

3.2.1.4. Le voisinage orthographique$^{(T)}$

Coltheart *et al.* (1977) ont suggéré que la densité du voisinage orthographique[9] pouvait influencer la lecture d'un mot. Relativement controversé, cet effet varie en fonction de la langue considérée, de la tâche à effectuer ou encore du fait que l'on considère la densité des voisins ou leur fréquence. Ainsi, Andrews (1997) rapporte un effet facilitateur pour l'anglais, lorsque la densité du voisinage augmente, mais un effet inhibiteur pour le français quand il existe au moins un voisin plus fréquent que le mot cible.

Cette hypothèse n'ayant jamais été explorée en lisibilité, nous avons modélisé l'effet de voisinage sous divers angles : le nombre de voisins, la présence d'un voisin plus fréquent, le nombre de voisins plus fréquents et les fréquences cumulées de tous les voisins. Ces variables ont été calculées sur la base de la liste de voisins distribuée dans le cadre du projet Lexique 3, qui comprend 128 919 formes fléchies. Pour chacune d'elles, nous avons considéré diverses statistiques : la moyenne, la médiane et les 75^{e} et 90^{e} percentiles, ce qui nous donne au final treize prédicteurs.

3.2.1.5. La longueur des mots$^{(C)}$

Très populaire en lisibilité, l'emploi de la longueur moyenne des mots remonterait aux travaux de Bear (1927, cité par Johnson, 1930). Son usage est justifié par l'hypothèse que plus un mot est long, plus il est difficile à décoder. Longtemps débattue, cette hypothèse a été confirmée par Vitu *et al.* (1990) et serait liée au fait qu'un mot plus long rend plus probable un repositionnement du regard lors de sa reconnaissance, ce qui allonge le temps de décodage.

En lisibilité, cet effet a été abordé soit *via* des mesures fondées sur les syllabes (Flesch, 1948), soit sur les lettres (Smith, 1961). Nous avons retenu la seconde approche et défini quatorze variables représentant soit une statistique (moyenne, médiane, percentile) soit un seuil (proportion de mots du texte de plus de n lettres).

3.2.2. *Les variables syntaxiques*

Le niveau syntaxique constitue une deuxième piste d'investigation traditionnelle en lisibilité, qui ressort non plus de l'étape de décodage des mots, mais déjà de celle de la compréhension du texte. Bien que généralement considérés comme moins efficaces que les variables lexicales, les prédicteurs syntaxiques ont fait la preuve qu'ils

9. Ces auteurs définissent comme voisins orthographiques deux mots qui ne se différencient que par une lettre, par exemple le mot SAGE a notamment comme voisins MAGE et SALE.

peuvent être combinés avec succès aux premières pour améliorer les performances des formules. Nous avons dès lors exploré trois sous-familles : la longueur des phrases, la complexité des formes verbales et les ratios de catégories grammaticales.

3.2.2.1. La longueur des phrases$^{(C)}$

C'est durant la période classique que la longueur moyenne des phrases, mesurée en nombre de mots, connaît un large succès avec Dale et Chall (1948) et Flesch (1948). Bien que des travaux postérieurs aient démontré l'absence de relation causale avec la difficulté (Schlesinger, 1968), ce type de prédicteur s'est avéré très efficace et simple à calculer. Nous avons dès lors défini douze variables sur la base de différentes statistiques représentant la distribution du nombre de mots par phrase dans les textes (moyenne, médiane, percentiles, etc.). L'une d'elles, le pourcentage de phrases de plus de trente mots, est inspirée des travaux de Daoust *et al.* (1996).

3.2.2.2. Les formes verbales$^{(T)}$

L'hypothèse que sous-tend cette famille est que certaines formes verbales (temps ou modes) sont plus ardues à déchiffrer que d'autres. Peu de travaux en lisibilité se sont intéressés aux verbes. Gillie (1957) introduit le nombre de verbes finis dans sa formule, tandis que Daoust *et al.* (1996) définissent un ratio de verbes conjugués sur le nombre de mots du texte. Plus récemment, Heilman *et al.* (2007) prennent en compte quatre aspects verbaux en anglais : le présent, le passé, le perfectif et le continu. François (2009a) est le premier à suggérer que le temps et le mode peuvent se révéler d'excellents prédicteurs de la difficulté des textes dans un contexte de FLE, où leur acquisition se fait généralement de façon progressive et programmée. Toutefois, il n'approche la question que d'une façon binaire : ses variables encodent la présence ou l'absence d'un temps particulier.

Repérant les temps et les modes dans les textes à l'aide du TreeTagger (Schmid, 1994), nous avons ici répliqué son approche, définissant ainsi onze variables binaires. Toutefois, nous avons également défini, pour ces mêmes temps et modes, onze variables continues, à savoir la proportion d'un temps particulier par rapport au nombre de verbes dans le texte. Nous avons également répliqué le ratio de Daoust *et al.* (1996).

3.2.2.3. Les variables catégorielles$^{(C)}$

Tout texte consiste en une suite d'unités qui appartiennent à des classes grammaticales particulières. L'emploi de variables catégorielles suppose que la fréquence relative d'apparition de certaines classes serait indicatrice de la difficulté des textes. Le premier à envisager de manière systématique ce type de variables est Bormuth (1966). Sur la base de deux ensembles de classes grammaticales, il définit plus de quatre-vingts ratios dont le meilleur est le nombre de pronoms sur le nombre de conjonctions ($r = 0{,}805$). Par la suite, plusieurs chercheurs reprennent cette idée sous diverses formes (Daoust *et al.*, 1996 ; Henry, 1975 ; Graesser *et al.*, 2004). Dans notre cas, nous avons simplement repris les trente-trois catégories du TreeTagger et les avons simplifiées pour obtenir neuf catégories plus générales : adverbes, noms communs,

noms propres, articles, pronoms, prépositions, verbes, adjectifs et conjonctions. Nous avons également constitué des supergroupes de catégories : la classe des noms (qui regroupe les noms communs et noms propres), les mots grammaticaux et les mots lexicaux. Pour chacune de ces douze classes, nous avons calculé (1) sa cote [10] par rapport à l'une des autres catégories, (2) sa cote par rapport à toutes les autres catégories et (3) sa proportion par rapport au nombre de mots du texte. Cela donne 156 variables catégorielles.

3.2.3. *Les variables sémantiques*

L'importance des variables sémantiques (y compris les aspects discursifs et cognitifs) a été particulièrement mise en avant par le paradigme structuro-cognitiviste, bien que ni Miller et Kintsch (1980), ni Kemper (1983) ne soient parvenus à démontrer la supériorité de ces prédicteurs sur les variables lexico-syntaxiques. Des travaux plus récents (Pitler et Nenkova, 2008 ; Feng *et al.*, 2010) ont également exploré cette dimension, sans obtenir de preuves plus probantes de cette supposée supériorité du niveau sémantique. Il est probable que l'importance de cette classe de facteurs, sans être négligeable, doit être relativisée par rapport aux affirmations structuro-cognitivistes. Afin d'apporter davantage de données sur ce point, nous avons implémenté trois sous-classes de prédicteurs, indicatives du niveau de personnalisation du texte, de la densité des idées qu'il contient et de sa cohésion.

3.2.3.1. Niveau de personnalisation du texte$^{(C)}$

Dale et Tyler (1934) ont suggéré que les textes rédigés dans un style plus informel sont plus simples à lire. En particulier, ils se sont focalisés sur l'usage des pronoms personnels, qui constituent un indice intéressant de cette dimension familière. En effet, Daoust *et al.* (1996) observent que le « tu », lié à un univers plus familier, est bien un indice de facilité. Dans notre étude, nous avons défini douze variables, mesurant la proportion (sur le nombre de mots du texte) ou la cote d'une des six classes suivantes : S2 ; S1+P1 ; S2 + P2 ; S3+P3 ; P2 ; S1 + P1 + S2 + P2 [11].

3.2.3.2. Densité des idées$^{(T)}$

Au-delà de la difficulté sémantique des mots, il arrive que l'incompréhension naisse de l'association de termes simples. Ce phénomène, qui échappe aux formules de lisibilité classiques, a cependant été abordé dès 1934 avec McClusky (1934) et sa variable « nombre d'idées pour 100 mots ». Plus récemment, Kintsch *et al.* (1975) ont développé un modèle psychologique de la compréhension des textes et ont montré que la densité conceptuelle influençait le temps de lecture. Nous avons utilisé une variable fondée sur l'implémentation du modèle de Kintsch pour le français, *Densidées* (Lee

10. Pour rappel, étant donné deux catégories *a* et *b*, la cote de *a* est $\frac{p(a)}{p(b)}$, c'est-à-dire le ratio entre la probabilité de *a* et *b*.

11. S1 signifie les pronoms personnels de la 1re personne du singulier ; P1 inclut les pronoms personnels pluriels de la 1re personne, etc.

et al., 2010). Elle correspond au nombre de propositions dans le texte divisé par le nombre de mots.

3.2.3.3. Cohésion du texte$^{(T)}$

La cohésion textuelle a été explorée par de nombreuses études, aussi bien sur le plan macrostructurel (Carrell, 1984 ; Kintsch *et al.*, 1977) qu'interphrastique (Myers *et al.*, 1987). Bien qu'il reste des divergences entre leurs conclusions, ces études reconnaissent l'effet de la cohésion sur la vitesse de lecture. Dans cette étude, nous avons repris une technique automatique de mesure de la cohésion interphrastique introduite par Foltz *et al.* (1998) : il s'agit de calculer, pour un texte, le cosinus moyen de toutes les paires de phrases adjacentes. Chaque phrase est transformée en un sac de mots, dont la fréquence absolue est encodée dans un vecteur numérique. Ensuite, nous avons projeté ce vecteur dans un espace vectoriel à l'aide de deux méthodes : le *tf-idf* (*term frequency-inverse document frequency*), qui accorde plus d'importance aux mots spécifiques à un texte particulier, et l'analyse sémantique latente (LSA), qui réduit l'espace en créant des « clusters » de mots sémantiquement cohérents [12]. Sur la base de ces deux espaces, nous avons défini quatre variables, selon que l'on considère les lemmes ou les formes fléchies.

3.2.4. *Les variables FLE*

Cette dernière famille de variables cherche à prendre en compte les spécificités du contexte FLE. À l'exception de l'effet des congénères introduit précédemment, ce type de variables n'a guère été exploré, probablement parce que cela requiert de disposer de données d'entraînement spécialisées, ainsi que de définir un modèle spécifique à chaque paire de langues considérée. Dans cette étude, nous avons envisagé deux prédicteurs qui affectent la lecture en L2, mais ne sont pas spécifiques à une L1 particulière : les unités polylexicales et le genre des textes.

3.2.4.1. Les unités polylexicales$^{(T)}$

Les unités polylexicales (UPs) sont connues pour poser des problèmes aux apprenants de L2 lors de leur production (Bahns et Eldaw, 1993). En revanche, l'effet de ce type d'unités lexicales sur la lecture est moins bien connu, en particulier chez les débutants. Ozasa *et al.* (2007) ont utilisé la moyenne de la fréquence absolue des UPs dans un texte comme prédicteur, mais celle-ci est apparue non significativement corrélée à la difficulté. Dans une expérience postérieure, François et Watrin (2011) ont défini vingt-deux variables capturant diverses caractéristiques des UPs et ont détecté une corrélation significative entre plusieurs de ces prédicteurs et la difficulté de leurs

12. Nos expérimentations, par validation croisée, portant sur le nombre de dimensions à conserver pour la LSA, ont produit un résultat étonnant. L'espace final ne comporte que quinze dimensions, alors que l'on considère qu'un espace de trois cents dimensions représente généralement une heuristique satisfaisante.

textes. Nous avons repris ici ces mêmes variables, fondées sur la fréquence des UPs [13], leur structure syntaxique (ex. NOM NOM ou NOM ADJ, etc.), leur nombre dans le texte ou leur longueur moyenne en nombre de mots.

3.2.4.2. Le genre des textes$^{(C)}$

Finalement, nous avons défini cinq prédicteurs visant à reconnaître les dialogues. Ceux-ci s'inspirent de Henry (1975) et utilisent la présence de guillemets, le type de ponctuation, etc. Ce type de variables est notamment justifié par le fréquent emploi de dialogues dans les premiers stades de l'apprentissage d'une langue étrangère. De plus, les dialogues contiennent généralement des termes plus simples et traitent de sujets plus quotidiens que d'autres genres de textes (Flesch, 1948).

3.2.5. *Conclusion*

Au final, ce ne sont pas moins de 406 variables appartenant à quatre grandes familles qui ont été implémentées dans le cadre de cette étude : 170 lexicales, 191 syntaxiques, 18 sémantiques – qui représentent la dimension la plus difficile à automatiser –, et 27 aspects spécifiques à la lecture en L2. Nous proposons ainsi une comparaison relativement exhaustive des prédicteurs traditionnels, mais suggérons également plusieurs nouvelles variables dans le domaine : le nombre de voisins orthographiques, le TTR normalisé, la plupart des variables fondées sur les UPs, la densité des idées calculée *via Densidées*, le nombre d'absents du vocabulaire d'*Alter Ego* et la proportion des temps verbaux dans le texte.

3.3. *Les algorithmes d'apprentissage*

La dernière étape dans la conception de notre formule consiste à sélectionner le meilleur sous-ensemble de prédicteurs et à combiner ceux-ci au sein d'un algorithme de classification. Nous avons testé les six algorithmes suivants : la régression logistique ordinale et multinomiale (respectivement RLO et RLM) (Agresti, 2002), les arbres de classification (Breiman *et al.*, 1984), le *bagging* (Breiman, 1996) et le *boosting* (Schapire et Freund, 1997) (qui reposent tout deux sur des arbres de décision) et les machines à vecteurs de support (SVM) (Boser *et al.*, 1992). Étant donné que ces modèles sont largement utilisés dans la littérature, nous ne les détaillons pas ici et renvoyons le lecteur aux références susmentionnées.

4. Résultats

Sur la base de cette méthodologie, une série d'expériences ont été réalisées en deux étapes. Dans un premier temps, nous avons évalué indépendamment l'efficacité

13. Les fréquences utilisées pour cette famille sont les mêmes que celles employées pour les n-grammes (voir section 3.2.1).

de chaque variable comme prédicteur de la difficulté des textes de notre corpus. Ensuite, différents ensembles de prédicteurs ont été comparés au sein des six algorithmes précités.

4.1. *Analyse des prédicteurs*

Traditionnellement, l'efficacité d'une variable est jugée à l'aune de sa corrélation linéaire (r de Pearson) avec la difficulté des textes [14]. Cependant, plusieurs chercheurs (Bormuth, 1966 ; Dale et Chall, 1948) ont remis en question le présupposé que la relation entre un prédicteur et la difficulté est toujours linéaire. C'est pourquoi nous avons opté pour une approche qui vérifie d'abord la nature de cette relation à l'aide du test F de linéarité de Guilford (1965, 314) [15]. Ensuite, l'intensité de cette relation a été évaluée *via* une corrélation de Pearson (r) dans le cas linéaire ou *via* une corrélation de Spearman (ρ) dans le cas d'une relation monotonique croissante. Enfin, la normalité des variables a également été évaluée à l'aide du test de normalité de Shapiro-Wilk. Ces quatre statistiques ont été estimées sur dix échantillons d'environ 600 textes, obtenus par rééchantillonnage – avec remise – des 1 852 textes du corpus. Cette technique de rééchantillonnage a permis de diviser la variabilité des mesures d'évaluation par la racine du nombre d'échantillons (à savoir $\sqrt{10}$), en accord avec le théorème centrale limite.

Il n'est pas possible de rapporter ici l'ensemble des résultats de cette première étape. Le tableau 2 présente, pour un nombre restreint de prédicteurs, la p-valeur pour le test F de linéarité et, en fonction de ce résultat, la corrélation à utiliser (r ou ρ). Les variables ont été classées par familles et en fonction de la force de leur corrélation. En revanche, nous n'avons pas inclus les résultats pour les tests de normalité des variables, puisque tous les tests ont rejeté l'hypothèse de normalité, dans la plupart des cas avec une p-valeur inférieure à $0{,}001$.

Lorsque l'on compare les familles de variables, la supériorité des prédicteurs lexicaux est confirmée, même si l'on observe également des corrélations relativement élevées au sein des familles syntaxique (ex. NMP) ou sémantique (ex. avLocalLsa_Lem). Les variables spécifiques au FLE apparaissent moins performantes, bien que BINGUI obtienne un ρ intéressant.

14. Signalons que pour l'étape d'analyse corrélationnelle, il importe de sélectionner une échelle de mesure à adopter pour représenter la difficulté des textes. Il nous a semblé que l'échelle ordinale était la plus appropriée dans ce contexte. Il s'agit de l'option habituellement retenue par les institutions d'enseignement des langues qui organisent le parcours d'apprentissage en G classes de difficulté croissante.

15. Ce test compare la différence entre r^2 et η^2. Le premier fait en effet l'hypothèse que la relation entre les deux variables est linéaire, tandis que le second modélise cette relation de façon curvilinéaire. Lorsque la différence entre r^2 et η^2 est significative, cela veut dire que la relation entre les deux variables n'est pas linéaire.

Label	Description de la variable	r	ρ	$F(p)$
Prédicteurs lexicaux				
PA_Alterego1a	Proportion d'absents de la liste de *Alter Ego 1*	/	$0{,}65^{**}$	$< 0{,}001$
X90FFFC	90e percentile de la probabilité des mots de contenu (formes fléchies)	/	$-0{,}64^{**}$	$< 0{,}001$
PAGoug_2000	Proportion d'absents de la liste de Gougenheim (2 000 premiers mots)	/	$0{,}60^{**}$	$0{,}017$
ML3	Modèle unigramme (formes fléchies)	/	$-0{,}55^{**}$	$< 0{,}001$
NL90P	90e percentile du nombre de lettres par mot	/	$0{,}52^{**}$	$0{,}022$
NLM	Nombre moyen de lettres par mot	$0{,}48^{**}$	/	$0{,}084$
meanNGProb.G	Moyenne des probabilités des bigrammes de Google	$0{,}38^{**}$	/	$0{,}05$
TTR	*Type token ratio* (calculé sur les lemmes)	/	$-0{,}26^{**}$	$0{,}009$
MedNeigh+Freq	Médiane du nombre de voisins plus fréquents	$-0{,}22^{**}$	/	$0{,}359$
Prédicteurs syntaxiques				
NMP	Nombre moyen de mots par phrase	/	$0{,}62^{**}$	$0{,}014$
LSDaoust	Pourcentage de phrases de plus de 30 mots	/	$0{,}56^{**}$	$0{,}008$
PPres	Présence de participes présents	/	$0{,}44^{**}$	$0{,}003$
PPasse	Présence de participes passés	/	$0{,}39^{**}$	$< 0{,}001$
Pres_C	Proportion des verbes à l'indicatif présent	/	$-0{,}34^{**}$	$< 0{,}001$
PRO.PRE	Rapport des pronoms aux prépositions	$-0{,}18^{**}$	/	$0{,}226$
Prédicteurs sémantiques				
avLocalLsa_Lem	Cohésion interphrastique moyenne mesurée *via* une LSA	/	$0{,}63^{**}$	$0{,}01$
PP1P2	Proportion de pronoms personnels de la 1re et de la 2e pers. sur le nombre de mots	/	$-0{,}33^{**}$	$0{,}008$
Prédicteurs spécifiques au FLE				
BINGUI	Présence de guillemets de dialogue	/	$0{,}46^{**}$	$0{,}018$
NAColl	Proportion d'UPs nominales de structure NA	/	$0{,}29^{**}$	/

Tableau 2. *Corrélations de Pearson et de Spearman, et p-valeur pour le test F de linéarité. Le niveau de significativité des prédicteurs est indiqué de la manière suivante :* * *:* $< 0{,}05$ *et* ** *:* $< 0{,}001$.

En ce qui concerne l'efficacité des prédicteurs, il est étonnant de noter que les deux meilleurs indices lexico-syntaxiques sont des variables classiques : le nombre d'absents d'une liste de mots facile et le nombre moyen de mots par phrase. Toutefois, on note que plusieurs « variables TAL » semblent efficaces : la cohésion mesurée à l'aide de la LSA, le modèle unigramme, les informations sur les temps verbaux, etc. Nous reviendrons cependant sur cette question par la suite.

Enfin, notons que la majorité des variables ne sont pas distribuées selon une gaussienne et qu'elles sont rarement corrélées linéairement à la difficulté. Bien que déjà signalé par quelques auteurs, cette propriété mérite d'être mise en évidence. En effet, cette non-linéarité tend à invalider l'usage de modèles traditionnels tels que la régression linéaire en lisibilité.

4.2. *Comparaison des modèles*

Une fois identifiés les meilleurs prédicteurs, nous avons cherché à les combiner au sein d'un modèle de lisibilité. Pour ce faire, nous avons tout d'abord sélectionné, parmi nos 406 prédicteurs, quelques sous-ensembles de variables susceptibles de modéliser

efficacement la difficulté des textes du corpus. Ensuite, chacun de ces sous-ensembles a été utilisé au sein d'un des six algorithmes décrits à la section 3.3. Les détails de ces deux opérations sont rapportés à la section 4.2.1.

Cependant, préalablement à ces comparaisons, il a été nécessaire de rééchantillonner notre corpus, car des expérimentations préalables ont confirmé que le nombre de textes par classe doit être constant pour éviter de biaiser le modèle. La classe la moins peuplée, C2, comportant 108 textes, il n'a été possible d'utiliser qu'un total de 648 documents pour nos tests. Ces données ont ensuite été divisées en deux sous-corpus. 240 textes forment le corpus de développement sur lequel a été opérée la sélection automatique des variables et l'estimation des métaparamètres des différents algorithmes statistiques, à l'aide d'une *grid search*. Les 408 textes restants ont servi pour entraîner et évaluer les performances des modèles ainsi définis.

4.2.1. *Sélection des variables*

Étant donné le nombre élevé de prédicteurs (406) au regard de la quantité de données d'entraînement disponibles et leur dépendance informationnelle, il a été nécessaire de sélectionner un sous-ensemble de variables. Pour ce faire, nous avons recouru à trois techniques, qui incarnent chacune une vision différente de la lisibilité. Tout d'abord, nous voulions disposer d'un modèle comparable aux formules traditionnelles, qui servirait de *baseline* à l'étude. Comme il n'existait pas de formule disponible pour le FLE, nous avons défini notre propre modèle fondé sur deux variables classiques emblématiques : le nombre moyen de lettres par mot (NLM) et le nombre moyen de mots par phrase (NMP) [16]. La seconde approche repose, quant à elle, sur un postulat structuro-cognitiviste selon lequel des formules de lisibilité ne devraient pas se contenter d'inclure des prédicteurs structuro-syntaxiques, mais également sémantiques, cognitifs, etc. Il s'agit en fait de maximiser le type d'information disponible pour le modèle. Nous avons appelé cette approche « experte » et l'avons déclinée en deux versions. La première consiste à retenir la meilleure variable dans chacune des quatre familles (*Exp1*), tandis que la seconde élargit la sélection aux deux meilleurs prédicteurs par famille (*Exp2*), pour autant qu'ils soient dans deux sous-familles différentes.

Enfin, nous avons appliqué des techniques standard de sélection automatique de variables, qui envisagent la lisibilité comme un problème de classification automatique. Toutefois, comme certaines sous-familles de variables correspondent en réalité à la variation d'un même paramètre (ex. la taille k de la liste de Gougenheim, où $k \in \{1\,000, 2\,000, ..., 8\,000, 8\,875\}$), nous n'avons retenu dans ce cas que le prédicteur correspondant à la valeur du paramètre produisant la meilleure corrélation avec le critère (ex. PAGoug_2000 pour $k = 2\,000$), d'après les résultats de l'analyse corrélationnelle (voir section 4.1). L'objectif de cette présélection était de limiter les problèmes de mul-

16. Soulignons que le fait d'entraîner notre *baseline* sur le même corpus que les modèles à évaluer devrait théoriquement la rendre plus compétitive qu'un modèle développé dans des conditions différentes, telle la formule de Kandel et Moles (1958) (voir section 4.2.3).

Méthode	Classifieur	Ensemble des variables
Baseline	Tous les 6	NLM + NMP
Exp1	Tous les 6	PA-Alterego + NMP + avLocalLsa-Lem + BINGUI
Exp2	Tous les 6	PA-Alterego + X90FFFC + NMP + PPres + avLocalLsa-Lem + PP1P2 + BINGUI + NAColl
Auto	RLO	PA-Alterego + NMP + PPres + ML3
Auto	RLM	PA-Alterego + Cond + Imperatif + Impf + PPasse + PPres + Subi + Subp + BINGUI + TTR + NWS90 + LSDaoust + MedNeigh+Freq
Auto	SVM, arbres, *boosting, bagging*	46 variables

Tableau 3. *Sous-ensembles de variables obtenus au terme de la sélection des variables*

ticollinéarité. Au terme de cette étape, nous avons obtenu un ensemble de quarante-six variables [17]. Ensuite, nous avons appliqué une procédure de sélection automatique de variables de type pas à pas pour la régression logistique [18] et une régularisation pour les SVM. Dans le cas des arbres de décision, la sélection des variables est interne à l'algorithme. Le tableau 3 présente les sous-ensembles qui ont été obtenus à l'aide de ces trois méthodes.

4.2.2. *Évaluation des modèles*

Chacun de ces quatre sous-ensembles (*baseline*, *Exp1*, *Exp2*, ou *Auto*) a été testés au sein de l'un des six algorithmes statistiques retenus à la section 3.3. Nous avons dès lors dû mettre en place la méthodologie de comparaison suivante pour faire le tri parmi ces nombreux modèles. Tout d'abord, les performances de chaque modèle ont été évaluées à l'aide de cinq mesures d'évaluation : le coefficient de corrélation multiple (R), l'exactitude (acc), l'exactitude contiguë [19] ($acc - cont$), l'erreur type ($rmse$) et l'erreur moyenne absolue (mae). [20] Le choix de calculer ces cinq mesures a été guidé par la variété des pratiques dans la littérature et le désir de rendre nos résultats comparables à ces autres approches. Cependant, pour tout ce qui touche à l'optimisation des métaparamètres et la sélection automatisée des variables, nous n'avons utilisé que l'exactitude contiguë. En effet, cette mesure se focalise davantage sur les erreurs graves (de plus d'un niveau). Or, il nous a semblé plus important, dans un cadre éducatif, d'éviter, autant que possible, de faire des prédictions très erronées, plutôt que de maximiser l'exactitude des prédictions. Signalons qu'il est évidemment possible, en lien avec un cadre applicatif différent, d'envisager de plutôt optimiser les modèles en

17. Pour la liste complète des variables retenues, consulter (François, 2011b, 460).

18. Afin de réduire la variabilité de cette technique, la sélection a été répétée cent fois sur des échantillons obtenus par *bootstrapping .632* (Tufféry, 2007, 396-371) et seules les variables retenues au moins cinquante fois ont été considérées.

19. Elle est définie par Heilman *et al.* (2008) comme « la proportion de prédictions situées à un niveau de difficulté ou moins du niveau de référence ». Autrement dit, pour un texte de niveau n, les prédictions $n-1$, n ou $n+1$ seront considérées comme correctes.

20. À l'exception de l'exactitude contiguë, ces mesures sont bien connues et sont discutées dans les nombreuses introductions à la statistique, telles que (Howell, 2008).

Modèle	Classifieur	Métaparamètres	R	acc	acc − cont	rmse	mae
Hasard	/	/	/	16,6	44,4	/	/
Baseline	RLM	/	0,58	**35**	65,2	1,69	1,1
	RLO	/	0,59	32,5	65,9	1,61	1,08
	Boosting	iter = 40	0,58	32,9	**68,5**	1,59	1,07
	SVM	$\gamma = 0{,}05; C = 25$	0,62	**34,0**	**68,2**	1,51	1,06
Exp1	**RLM**	/	0,70	**39,4**	74,2	1,34	0,97
	RLO	/	0,72	36,8	**77,8**	1,25	0,95
	Boosting	iter = 100	0,66	34	74	1,43	1,01
	SVM	$\gamma = 0{,}006; C = 1$	0,70	35	72,1	1,41	1,01
Exp2	RLM	/	0,73	**41**	77	1,31	0,95
	RLO	/	0,74	40	76,6	1,25	0,94
	Boosting	iter = 40	0,70	38	76	1,33	0,97
	SVM	$\gamma = 0{,}002; C = 75$	0,73	**41**	**78**	1,28	0,94
Auto	RLM	/	0,71	44	75,7	1,32	0,95
	RLO	/	0,71	39,6	76,1	1,33	0,96
	Boosting	iter = 40	0,71	46	76,6	1,33	0,97
	SVM	$\gamma = 0{,}004; C = 5$	0,73	**49**	**79,6**	1,27	0,90

Tableau 4. *Performances des modèles au sein de chaque sous-ensemble de prédicteurs (*baseline, Exp1, Exp2 *et* Auto*), mis en correspondance avec les résultats obtenus par un modèle fondé uniquement sur le hasard*

fonction de l'exactitude. Cependant, ces deux mesures étant relativement consistantes, il n'est pas certain que cette démarche modifie beaucoup les résultats obtenus.

Ces cinq mesures d'évaluation ont été estimées pour chaque modèle à l'aide d'une procédure de validation croisée à dix échantillons, pratiquée selon un échantillonnage systématique, [21] ce qui a permis de comparer les modèles entre eux à l'aide d'un test-T pairé (Witten et Frank, 2005, 153-157). En ce qui concerne la comparaison des modèles, afin d'éviter d'effectuer un grand nombre de comparaisons, ce qui requiert d'adapter l'α des tests de significativité, nous avons décidé de nous limiter à trois comparaisons planifiées (Howell, 2008, 359) au sein de chaque sous-ensemble de variables. En effet, comme nos expériences préliminaires avaient démontré la supériorité de la régression logistique (RLO et RLM) et des SVM, nous avons uniquement comparé les performances de ces trois modèles au sein de chaque sous-ensemble, afin de retenir le meilleur modèle d'entre eux. Ensuite, les meilleurs modèles de chaque sous-ensemble de prédicteurs ont été comparés entre eux, afin de retenir le classifieur le plus performant comme notre modèle de lisibilité final. Notons que la comparaison a cette fois porté aussi bien sur l'exactitude que sur l'exactitude contiguë, afin de dégager le modèle le plus performant possible dans ces deux dimensions.

Les résultats de la plupart des modèles sont repris dans le tableau 4. Nous en avons omis certains pour des raisons de place, en particulier ceux reposant sur le *bagging* et les arbres, car leurs performances étaient toujours significativement inférieures aux autres algorithmes. Pour le premier sous-ensemble, c'est-à-dire la *baseline* « clas-

21. L'échantillonnage systématique au sein d'une validation croisée à k échantillons produit pour chaque modèle exactement les mêmes k échantillons, ce qui permet de contrôler l'erreur d'échantillonnage.

sique », on observe que la régression logistique multinomiale (RLM) produit les prédictions les plus exactes. Elle est suivie à ce niveau par les SVM, tandis que l'exactitude contiguë est supérieure pour le *boosting* et les SVM. C'est ce dernier modèle qui a été retenu, car ni son acc, ni son $acc - cont$ ne sont significativement différentes de celles de la RLM et il obtient les meilleures performances combinées en termes d'acc et $acc - cont$.

Au niveau des modèles experts, il est intéressant de noter que pour le sous-ensemble *Exp1*, ce sont les modèles logistiques qui se démarquent, même par rapport aux SVM : la RLM est significativement plus exacte que les SVM ($t(9) = 2{,}34$; $p = 0{,}02$), tandis que la RLO réduit significativement le nombre d'erreurs graves par rapport à ce même modèle ($t(9) = 3{,}18$; $p = 0{,}006$). Par conséquent, il convient de retenir l'un des deux modèles de régression pour *Exp1*, en fonction de la mesure privilégiée. Par ailleurs, on note que l'ajout d'une variable sémantique et d'une spécifique au FLE produit déjà un modèle plus exact que la *baseline*, à un niveau proche de la significativité ($t(9) = 1{,}77$; $p = 0{,}055$).

Pour *Exp2*, les trois modèles (SVM, RLM, et RLO) obtiennent des performances très similaires. Par conséquent, ces trois modèles apparaissent équivalents et nous n'avons retenu les SVM pour la suite des comparaisons que parce qu'elles obtiennent des scores très légèrement supérieurs. Les SVM, comme les modèles logistiques, se révèlent significativement supérieures à la *baseline* ($t(9) = 2{,}36$; $p = 0{,}02$ pour l'acc). Il est intéressant de noter que, en revanche, l'apport de quatre variables supplémentaires dans le modèle ne permet pas de dépasser significativement les performances d'*Exp1* : la différence d'exactitude est non significative ($t(9) = 0{,}66$; $p = 0{,}26$), alors que la différence d'exactitude contiguë est à la limite de la significativité ($t(9) = 1,76$; $p = 0.056$).

Afin de mieux appréhender la participation de chacune de ces huit variables au modèle *Exp2*, nous avons entraîné huit modèles par SVM comprenant $p - 1$ variables et les avons comparés, en termes d'acc et d'$acc - cont$, au modèle complet. Le gain marginal de chacune de ces variables est repris à la figure 1. On y voit que les deux variables les plus informatives sont le nombre d'absents de la liste d'*Alter Ego* et l'indicateur de dialogue (BINGUI).

Enfin, les meilleurs modèles – et de loin – sont obtenus *via* la procédure de sélection automatique de variables, dont le meilleur d'entre eux est clairement celui fondé sur les SVM. Ce dernier modèle est effectivement significativement plus précis que la RLM ($p = 0{,}02$) et la RLO ($p = 0{,}05$) et son exactitude contiguë frôle la significativité dans les deux cas ($p = 0{,}08$ et $p = 0{,}07$). Par ailleurs, son exactitude augmente de $8\ \%$ par rapport au meilleur modèle expert (*Exp2*), ce qui est tout à fait significatif ($t(9) = 2{,}61$; $p = 0{,}01$). C'est donc le modèle que nous avons retenu pour notre modèle de lisibilité.

Nous avons ensuite évalué la contribution spécifique de chaque famille de prédicteurs au sein du modèle SVM des deux façons suivantes : d'une part, nous avons entraîné des modèles SVM contenant les prédicteurs d'une seule famille ; d'autre part,

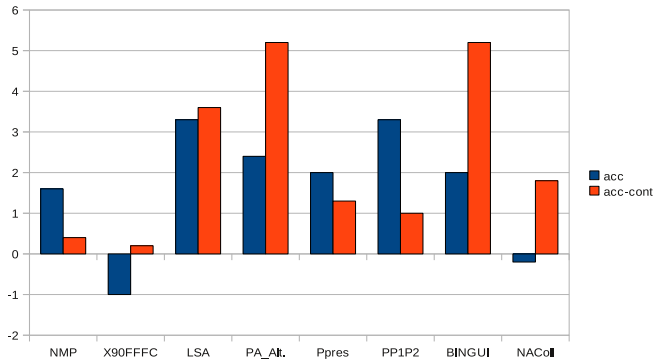

Figure 1. *Apport marginal (en pourcentage) de chacune des variables dans le modèle Exp2*

nous avons comparé le même type de modèle, mais contenant des variables issues des trois autres familles. Les résultats de ces expériences sont rapportés au tableau 5. Il en ressort que les variables lexicales sont bien les plus exactes ($acc = 40{,}5\ \%$), de même que celles qui permettent de réduire le mieux les erreurs graves ($acc - cont = 75{,}6\ \%$). De plus, il s'agit du seul ensemble de variables dont l'absence dans le modèle influence significativement l'exactitude contiguë. Les aspects syntaxiques se révèlent également efficaces, avec une $acc = 39{,}3\ \%$, mais une $acc - cont$ légèrement plus faible. En revanche, les prédicteurs sémantiques et spécifiques au FLE apparaissent clairement moins performants, même s'ils améliorent tout de même légèrement l'exactitude totale du modèle.

	Famille uniquement		**Tous sauf famille**	
	acc	$acc - cont$	acc	$acc - cont$
Lexical	40,5	75,6	41,1	73,5
Syntaxique	39,3	69,5	43,2	78,4
Sémantique	28,8	61,5	47,8	79,2
FLE	24,9	58,5	47,8	79,6

Tableau 5. *Exactitude et exactitude contiguë (en pourcentage) pour les modèles employant soit une seule famille de prédicteurs, soit les quarante-six prédicteurs excepté ceux de cette famille*

Afin de mieux comprendre l'effet des différentes familles de variables sur la qualité des prédictions, l'$acc - cont$ obtenue pour chacun des six niveaux a été calculée, non seulement pour le modèle complet, mais également pour chacun des quatre modèles ci-dessus, entraînés sur une seule famille de variables. Le tableau 6 rapporte les exactitudes contiguës moyennes calculées sur les dix échantillons. De manière générale,

on note que les prédictions sont meilleures pour les niveaux faciles (A1 et A2), où l'on rencontre une langue plus typée : les phrases sont courtes, les mots simples et les structures relativement redondantes. À l'inverse, les niveaux moyens (B1 et B2) sont ceux où le modèle éprouve le plus de mal à discriminer correctement les textes.

	A1	A2	B1	B2	C1	C2
Lexical	86	75	65	69	83	75
Syntaxique	94	83	53	44	76	67
Sémantique	65	68	49	26	85	81
FLE	58	67	27	43	80	79
Modèle complet	95	88	79	64	83	69

Tableau 6. *Exactitude contiguë moyenne par niveau (en pourcentage), calculée sur les dix échantillons*

Lorsque l'on analyse les performances par familles de variables, il apparaît que les variables syntaxiques discriminent mieux les textes plus simples, tandis que les variables sémantiques – telles que le niveau de cohérence – sont plus utiles pour discriminer les textes complexes. De même, les variables spécifiques au FLE effectuent de meilleures prédictions pour les textes difficiles, probablement parce qu'elles concernent des caractéristiques de plus haut niveau (les unités polylexicales, le genre de texte, etc.). Enfin, les prédicteurs lexicaux se révèlent les plus robustes, discriminant efficacement les textes de tout niveau. Ces résultats apparaissent relativement cohérents avec la littérature et les modèles du processus de lecture. Chall et Dale (1995, 111) expliquent ainsi que les mesures classiques (c'est-à-dire celles qui concernent surtout les aspects lexicaux et syntaxiques) semblent plus efficaces pour discriminer les niveaux de difficulté plus bas, tandis que les facteurs structuro-cognitifs sont meilleurs pour les textes de niveau plus avancé.

4.2.3. *Comparaison avec la littérature*

Les comparaisons avec les rares autres modèles disponibles pour le FLE sont assez délicates, à cause des différences de populations cibles, de types de corpus, etc. Par exemple, si Uitdenbogerd (2005) rapporte une corrélation multiple supérieure à la nôtre ($0{,}87$ contre $0{,}73$), son modèle se focalise uniquement sur un type de L2 (lecteur anglophone) et un type de texte (narratif), sans compter que le R semble obtenu sur le corpus d'entraînement.

C'est pourquoi, nous n'avons pu strictement comparer nos résultats qu'avec les deux modèles précédents. Le premier d'entre eux est une formule de lisibilité classique pour le français L1 développée par Kandel et Moles (1958) et qui constitue une adaptation de la formule de Flesh. Bien qu'elle n'ait pas été conçue spécifiquement pour le FLE, elle est l'une des formules les plus connues pour le français et repose sur deux variables génériques (nombre de syllabes et de mots), dont la capacité prédictive ne varierait guère dans un contexte L2 (Greenfield, 2004). Nous l'avons dès lors ap-

pliqué à notre propre corpus et avons obtenu des performances au mieux équivalentes à notre propre *baseline* : $R = 0{,}55$ et $acc = 33\ \%$.

Le second modèle testé est celui de François (2009a), utilisant une régression logistique et les dix variables suivantes : un modèle unigramme, le nombre moyen de lettres par mot, le nombre moyen de mots par phrase et sept variables binaires de temps. Sur nos données, ce modèle atteint une exactitude de $41\ \%$ et une exactitude contiguë de $72{,}7\ \%$. Notre nouvelle approche est donc $8\ \%$ plus exacte que cette formule précédente, ce qui est significatif ($t(9) = 3{,}72$; $p = 0{,}002$). Ces comparaisons confirment le gain obtenu par cette nouvelle approche.

Afin d'offrir au lecteur une impression plus générale des performances de notre modèle par rapport à d'autres approches en lisibilité computationnelle, le tableau 7 présente des résultats publiés dans des études récentes, portant essentiellement sur l'anglais L1. Les valeurs présentées doivent être interprétées avec prudence, car plus grand est le nombre de niveaux de difficulté utilisés dans ces modèles, plus la tâche de classification est ardue. C'est pourquoi Heilman *et al.* (2008) obtiennent une $acc - cont$ qui ne dépasse pas $52\ \%$ pour leur modèle à douze classes, tandis que notre propre modèle, à six classes, obtient $80\ \%$. Une façon plus fiable de comparer deux approches est de calculer le gain d'$acc - cont$ par rapport au hasard. Chez Schwarm et Ostendorf (2005), qui se basent aussi sur des SVM, ce gain est compris entre $24{,}5\ \%$ et $29\ \%$, tandis qu'il atteint en moyenne $36\ \%$ pour notre modèle. On voit donc que les performances de notre approche se situent bien dans la lignée des approches récentes pour l'anglais, mais également que prédire automatiquement la difficulté de textes constitue un problème complexe.

Étude	**♯ classes**	**Langue**	**Exac. cont.**	**R**	**RMSE**
Collins-Thompson et Callan (2004)	6	ENG.	/	0,64	/
Collins-Thompson et Callan (2004)	12	ENG.	/	0,79	/
Collins-Thompson et Callan (2004)	5	FR.	/	0,64	/
Schwarm et Ostendorf (2005)	4	ENG.	79% à 94,5%	/	/
Heilman *et al.* (2007)	12	ENG.	/	0,72	2,17
Heilman *et al.* (2007)	4	ENG. (L2)	/	0,81	0,66
Heilman *et al.* (2008)	12	ENG.	45%	0,58	2,94
Heilman *et al.* (2008)	12	ENG.	52%	0,77	2,24
Pitler et Nenkova (2008)	5	ENG.	/	0,78	/
Feng *et al.* (2009)	4	ENG.	/	− 0,34	0,57
Feng *et al.* (2010)	4	ENG.	/	/	/
Kate *et al.* (2010)	5	ENG.	/	0,82	/
Notre modèle	6	FR. (L2)	80%	0,73	1,23

Tableau 7. *Comparaison des performances de notre modèle par rapport aux principales études de la littérature*

5. Discussion et conclusions

Cet article propose une nouvelle formule de « lisibilité computationnelle » pour le FLE, capable d'effectuer des prédictions exprimées en fonction de l'échelle du

CECR, désormais largement utilisée dans l'enseignement du FLE. Ce modèle a été obtenu après comparaison de quatre techniques de sélection de variables et de six algorithmes statistiques. Notre étude a établi que le meilleur de ces modèles est un classifieur SVM reposant sur quarante-six variables représentatives de différents niveaux d'informations textuelles. Ce modèle atteint en effet une exactitude de $49\ \%$ et une exactitude contiguë de $80\ \%$, ce qui est largement meilleur que notre modèle précédent (François, 2009a). Plus concrètement, il se caractérise par une erreur moyenne absolue (mae) de $0{,}9$, ce qui signifie que le modèle se trompe, en moyenne, de moins d'un niveau par rapport au niveau de référence des textes évalués. Au-delà de ce résultat, cette étude propose également quelques nouveaux prédicteurs de la difficulté, inspirés des travaux sur le processus de lecture en L2, à savoir les variables fondées sur les voisins orthographiques, les unités polylexicales, le TTR normalisé et, dans une moindre mesure, les temps verbaux.

En ce qui concerne l'apport du TAL et de l'apprentissage automatisé à la lisibilité, le résultat est plus mitigé. À l'issue de nos expériences, les meilleures variables considérées indépendamment sont plutôt de nature « classique », à l'exception de celle qui repose sur la LSA. Cependant, cette dernière n'a pas été retenue lors de la sélection automatique des variables à cause d'une trop grande colinéarité avec les autres prédicteurs. À première vue, l'apport du TAL pourrait donc paraître peu convaincant. Pourtant, il semble avoir entraîné des gains significatifs de performances par rapport à une formule classique comme celle de Kandel et Moles (1958).

Afin de déterminer plus précisément l'apport du TAL et de l'apprentissage automatisé, nous avons effectué une seconde expérience, plus spécifique à cette question, rapportée en détail dans (François et Miltsakaki, 2012). Dans celle-ci, nous avons découpé nos quarante-six variables en deux sous-ensembles de même taille, le premier regroupant les variables reconnues comme « classiques » et le second les variables d'inspiration TAL (*cf.* les notations ^{T} et ^{C} de la section 3.2). Ensuite, nous avons entraîné un modèle SVM sur chacun de ces ensembles et avons comparé leurs performances selon la méthode décrite précédemment. Il est apparu que le modèle SVM combinant les variables TAL était $3{,}8\ \%$ plus exact, ce qui n'est pas suffisant pour être significatif ($t(9) = 1{,}50$; $p = 0{,}08$). Cependant, seules, les variables « classiques » ne dépassaient pas $37{,}5\ \%$, alors que lorsqu'elles étaient combinées avec les aspects TAL, le modèle atteignait $49\ \%$, comme nous l'avons vu plus haut. Les variables TAL semblent donc bien apporter une information supplémentaire par rapport aux prédicteurs classiques, même si l'on ne peut exclure un effet du nombre de variables.

En ce qui concerne l'apprentissage automatisé, l'idée est ici que des modèles plus complexes, comme les SVM, pourraient mieux tirer parti du plus grand nombre de prédicteurs, ainsi que de la relation non linéaire de certains d'entre eux avec la difficulté. Nous avons donc cette fois comparé les variables linéaires et celles non linéaires (d'après le test F) à l'aide tantôt d'une régression linéaire (RL) ou d'une SVM. Les résultats sont particulièrement troublants, puisque si les SVM améliorent bien l'exactitude du même ensemble de variables ($+\ 11\ \%$ pour les prédicteurs linéaires et $+\ 8\ \%$ pour les non linéaires), l'exactitude contiguë des deux modèles n'est pas significative-

ment différente ($+ 1\ \%$ pour les prédicteurs linéaires et $- 5\ \%$ pour les non linéaires). Il semble donc que le nombre d'erreurs graves soit le même et qu'une part de la supériorité des modèles automatisés pourrait s'expliquer par le type de mesure d'évaluation employé.

Notre étude, si elle a proposé une nouvelle approche pour la lisibilité du FLE, n'a pas véritablement démontré la supériorité du TAL en lisibilité. La question de l'apport du TAL en lisibilité reste donc ouverte, en particulier depuis l'étude de Nelson *et al.* (2012). Ces auteurs ont effectué une étude corrélationnelle entre six formules, dont une appartient au paradigme de la « lisibilité computationnelle », et cinq séries de textes de référence et ont observé les corrélations les plus faibles pour le modèle computationnel. Par ailleurs, il reste aussi la question de la complexité du modèle. Bien qu'étant légèrement moins exacts, certains de nos modèles fondés sur la régression logistique atteignent des résultats honorables avec des moyens moindres. Dans un contexte où la vitesse d'exécution importe – tel que la recherche de textes sur le Web –, il serait probablement nécessaire d'équilibrer ces deux dimensions, sachant que les variables classiques sont nettement plus rapides à calculer.

Parmi d'autres perspectives, il est clair que l'aspect spécifique au FLE de notre formule pourrait être accentué. Une question importante qui n'a guère été traitée dans la littérature est de déterminer si des prédicteurs qui prennent en compte des aspects particuliers du lecteur de L2 – en particulier sa L1 – permettent d'obtenir des performances supérieures à celles d'un modèle générique tel que le nôtre. Dans la même veine, on peut imaginer des formules de lisibilité personnalisables, qui s'adapteraient peu à peu à un utilisateur spécifique au fur et à mesure de leur utilisation. On voit donc qu'il reste bien des pistes à explorer dans le domaine afin d'améliorer les modèles existants.

Remerciements

Ce travail a été subventionné par une bourse F.N.R.S et la rédaction de cet article a été rédigé grâce au soutien d'une bourse de la B.A.E.F. Par ailleurs, nous remercions particulièrement Bernadette Dehottay, dont l'aide précieuse nous a permis de rassembler notre corpus de référence.

6. Bibliographie

Agresti A., *Categorical Data Analysis. 2nd edition*, Wiley-Interscience, New York, 2002.

Aluisio S., Specia L., Gasperin C., Scarton C., « Readability assessment for text simplification », *Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, Los Angeles, p. 1-9, 2010.

Andrews S., « The effect of orthographic similarity on lexical retrieval : Resolving neighborhood conflicts », *Psychonomic Bulletin & Review*, vol. 4, n° 4, p. 439-461, 1997.

Bahns J., Eldaw M., « Should We Teach EFL Students Collocations ? », *System*, vol. 21, n° 1, p. 101-14, 1993.

Berthet A., Hugot C., Kizirian V., Sampsonis B., Waendendries M., *Alter Ego 1*, Hachette, Paris, 2006a.

Berthet A., Hugot C., Kizirian V., Sampsonis B., Waendendries M., *Alter Ego 2*, Hachette, Paris, 2006b.

Bormuth J., « Readability : A new approach », *Reading research quarterly*, vol. 1, n° 3, p. 79-132, 1966.

Boser B., Guyon I., Vapnik V., « A training algorithm for optimal margin classifiers », *Proceedings of the fifth annual workshop on Computational learning theory*, p. 144-152, 1992.

Bowers J., « In defense of abstractionist theories of repetition priming and word identification », *Psychonomic bulletin and review*, vol. 7, n° 1, p. 83-99, 2000.

Breiman L., « Bagging predictors », *Machine learning*, vol. 24, n° 2, p. 123-140, 1996.

Breiman L., Friedman H., Olsen R., Stone J., *Classification and regression trees*, Chapman & Hall, New York, 1984.

Carrell P., « The effects of rhetorical organization on ESL readers », *TESOL quarterly*, vol. 18, n° 3, p. 441-469, 1984.

Chall J., Dale E., *Readability Revisited : The New Dale-Chall Readability Formula*, Brookline Books, Cambridge, 1995.

Chen S., Goodman J., « An empirical study of smoothing techniques for language modeling », *Computer Speech and Language*, vol. 13, n° 4, p. 359-393, 1999.

Collins-Thompson K., Callan J., « A language modeling approach to predicting reading difficulty », *Proceedings of HLT/NAACL 2004*, Boston, USA, p. 193-200, 2004.

Collins-Thompson K., Callan J., « Predicting reading difficulty with statistical language models », *Journal of the American Society for Information Science and Technology*, vol. 56, n° 13, p. 1448-1462, 2005.

Coltheart M., Davelaar E., Jonasson T., Besner D., « Access to the internal lexicon », *in* S. Dornic (ed.), *Attention and Performance VI*, Academic Press, London, p. 535-555, 1977.

Conseil de l'Europe ., *Cadre européen commun de référence pour les langues*, 2001.

Cornaire C., « La lisibilité : essai d'application de la formule courte d'Henry au français langue étrangère », *Canadian Modern Language Review*, vol. 44, n° 2, p. 261-273, 1988.

Dale E., Chall J., « A formula for predicting readability », *Educational research bulletin*, vol. 27, n° 1, p. 11-28, 1948.

Dale E., Tyler R., « A study of the factors influencing the difficulty of reading materials for adults of limited reading ability », *The Library Quarterly*, vol. 4, p. 384-412, 1934.

Daoust F., Laroche L., Ouellet L., « SATO-CALIBRAGE : Présentation d'un outil d'assistance au choix et à la rédaction de textes pour l'enseignement », *Revue québécoise de linguistique*, vol. 25, n° 1, p. 205-234, 1996.

Feng L., Elhadad N., Huenerfauth M., « Cognitively motivated features for readability assessment », *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, p. 229-237, 2009.

Feng L., Jansche M., Huenerfauth M., Elhadad N., « A Comparison of Features for Automatic Readability Assessment », *COLING 2010 : Poster Volume*, p. 276-284, 2010.

Flesch R., « A new readability yardstick », *Journal of Applied Psychology*, vol. 32, n° 3, p. 221-233, 1948.

Foltz P., Kintsch W., Landauer T., « The measurement of textual coherence with latent semantic analysis », *Discourse processes*, vol. 25, n° 2, p. 285-307, 1998.

François T., « Combining a statistical language model with logistic regression to predict the lexical and syntactic difficulty of texts for FFL », *Proceedings of the 12th Conference of the EACL : Student Research Workshop*, p. 19-27, 2009a.

François T., « Modèles statistiques pour l'estimation automatique de la difficulté de textes de FLE », *11e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, 2009b.

François T., « La lisibilité computationnelle : un renouveau pour la lisibilité du français langue première et seconde ? », *International Journal of Applied Linguistics (ITL)*, vol. 160, p. 75-99, 2011a.

François T., Les apports du traitement automatique du langage à la lisibilité du français langue étrangère, PhD thesis, Université Catholique de Louvain, 2011b. Thesis Supervisors : Cédrick Fairon and Anne Catherine Simon.

François T., Miltsakaki E., « Do NLP and machine learning improve traditional readability formulas ? », *Proceedings of the 2012 Workshop on Predicting and improving text readability for target reader populations (PITR2012)*, 2012.

François T., Watrin P., « On the contribution of MWE-based features to a readability formula for French as a foreign language », *Proceedings of the International Conference RANLP 2011*, 2011.

Fry E., « A readability formula that saves time », *Journal of reading*, vol. 11, n° 7, p. 513-578, 1968.

Gale W., Sampson G., « Good-Turing frequency estimation without tears », *Journal of Quantitative Linguistics*, vol. 2, n° 3, p. 217-237, 1995.

Galley M., McKeown K., « Improving word sense disambiguation in lexical chaining », *International Joint Conference on Artificial Intelligence*, vol. 18, p. 1486-1488, 2003.

Gillie P., « A simplified formula for measuring abstraction in writing. », *Journal of applied psychology*, vol. 41, n° 4, p. 214-217, 1957.

Gougenheim G., Michéa R., Rivenc P., Sauvageot A., *L'élaboration du français fondamental (1er degré)*, Didier, Paris, 1964.

Graesser A., McNamara D., Louwerse M., Cai Z., « Coh-Metrix : Analysis of text on cohesion and language », *Behavior Research Methods, Instruments, & Computers*, vol. 36, n° 2, p. 193-202, 2004.

Greenfield J., « Readability formulas for EFL », *Japan Association for Language Teaching*, vol. 26, n° 1, p. 5-24, 2004.

Guilford J., *Fundamental statistics in psychology and education*, McGraw-Hill, New-York, 1965.

Heilman M., Collins-Thompson K., Callan J., Eskenazi M., « Combining lexical and grammatical features to improve readability measures for first and second language texts », *Proceedings of NAACL HLT*, p. 460-467, 2007.

Heilman M., Collins-Thompson K., Eskenazi M., « An analysis of statistical models and features for reading difficulty prediction », *Association for Computational Linguistics*, vol. The

3rd Workshop on Innovative Use of NLP for Building Educational Applications, p. 1-8, 2008.

Henry G., *Comment mesurer la lisibilité*, Labor, 1975.

Howell D., *Méthodes statistiques en sciences humaines, 6e édition*, De Boeck, Bruxelles, 2008.

Howes D., Solomon R., « Visual duration threshold as a function of word probability », *Journal of Experimental Psychology*, vol. 41, no 40, p. 1-4, 1951.

Johnson G., « An objective method of determining reading difficulty », *The Journal of Educational Research*, vol. 21, no 4, p. 283-287, 1930.

Kandel L., Moles A., « Application de l'indice de Flesch à la langue française », *Cahiers Études de Radio-Télévision*, vol. 19, p. 253-274, 1958.

Kate R., Luo X., Patwardhan S., Franz M., Florian R., Mooney R., Roukos S., Welty C., « Learning to predict readability using diverse linguistic features », *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, p. 546-554, 2010.

Kemper S., « Measuring the inference load of a text. », *Journal of Educational Psychology*, vol. 75, no 3, p. 391-401, 1983.

Kincaid J., Fishburne R., Rodgers R., Chissom B., « Derivation of new readability formulas for navy enlisted personnel », *Research Branch Report*, 1975.

Kintsch W., Kozminsky E., Streby W., McKoon G., Keenan J., « Comprehension and recall of text as a function of content variables1 », *Journal of Verbal Learning and Verbal Behavior*, vol. 14, no 2, p. 196-214, 1975.

Kintsch W., Mandel T., Kozminsky E., « Summarizing scrambled stories. », *Memory & Cognition*, vol. 5, p. 547-552, 1977.

Kintsch W., Vipond D., « Reading comprehension and readability in educational practice and psychological theory », *in* L. Nilsson (ed.), *Perspectives on Memory Research*, Lawrence Erlbaum, Hillsdale, NJ, p. 329-365, 1979.

Koda K., *Insights into second language reading : A cross-linguistic approach*, Cambridge University Press, Cambridge, 2005.

Krashen S., « We acquire vocabulary and spelling by reading : Additional evidence for the input hypothesis », *The Modern Language Journal*, vol. 73, no 4, p. 440-464, 1989.

Lee H., Gambette P., Maillé E., Thuillier C., « Densidées : calcul automatique de la densité des idées dans un corpus oral », *Actes de la 12e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des langues (RECITAL)*, 2010.

Lively B., Pressey S., « A method for measuring the "vocabulary burden" of textbooks », *Educational Administration and Supervision*, vol. 9, p. 389-398, 1923.

Lorge I., « Predicting readability », *the Teachers College Record*, vol. 45, no 6, p. 404-419, 1944.

Lupker S., « Visual word recognition : Theories and findings », *in* M. Snowling, C. Hulme (eds), *The science of reading : A handbook*, Blackwell Publishing, Oxford, p. 39-60, 2005.

McClusky H., « A Quantitative Analysis of the Difficulty of Reading Materials », *The Journal of Educational Research*, vol. 28, p. 276-282, 1934.

Mesnager J., « La lisibilité dans la littérature enfantine », *Les actes de lecture*, vol. 13, p. 31-39, 1986.

Meyer B., « Reading research and the composition teacher : The importance of plans », *College composition and communication*, vol. 33, n° 1, p. 37-49, 1982.

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., « Quantitative analysis of culture using millions of digitized books », *Science*, vol. 331, n° 6014, p. 176-182, 2011.

Miller J., Kintsch W., « Readability and recall of short prose passages : A theoretical analysis », *Journal of Experimental Psychology : Human Learning and Memory*, vol. 6, n° 4, p. 335-354, 1980.

Monsell S., « The nature and locus of word frequency effects in reading », *in* D. Besner, G. Humphreys (eds), *Basic processes in reading : Visual word recognition*, Lawrence Erlbaum Associates Inc., Hillsdale, NJ, p. 148-197, 1991.

Myers J., Shinjo M., Duffy S., « Degree of causal relatedness and memory », *Journal of Memory and Language*, vol. 26, n° 4, p. 453-465, 1987.

Nelson J., Perfetti C., Liben D., Liben M., Measures of text difficulty : Testing their predictive value for grade levels and student performance, Technical report, Technical report, The Council of Chief State School Officers, 2012.

New B., Brysbaert M., Veronis J., Pallier C., « The use of film subtitles to estimate word frequencies », *Applied Psycholinguistics*, vol. 28, n° 04, p. 661-677, 2007.

O'Connor R., Bell K., Harty K., Larkin L., Sackor S., Zigmond N., « Teaching reading to poor readers in the intermediate grades : A comparison of text difficulty », *Journal of Educational Psychology*, vol. 94, n° 3, p. 474-485, 2002.

Ozasa T., Weir G., Fukui M., « Measuring readability for Japanese learners of English », *Proceedings of the 12th Conference of Pan-Pacific Association of Applied Linguistics*, 2007.

Pitler E., Nenkova A., « Revisiting readability : A unified framework for predicting text quality », *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 186-195, 2008.

Prasad R., Dinesh N., Lee A., Miltsakaki E., Robaldo L., Joshi A., Webber B., « The penn discourse treebank 2.0 », *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, p. 2961-2968, 2008.

Redish J., Selzer J., « The place of readability formulas in technical communication », *Technical communication*, vol. 32, n° 4, p. 46-52, 1985.

Reitsma P., « Reading practice for beginners : Effects of guided reading, reading-while-listening, and independent reading with computer-based speech feedback », *Reading Research Quarterly*, vol. 23, n° 2, p. 219-235, 1988.

Schapire R., Freund Y., « A decision theoretic generalization of on-line learning and an application to boosting », *Journal Computer and System Sciences*, vol. 55, p. 119-139, 1997.

Schlesinger I., *Sentence structure and the reading process*, Mouton, The Hague, 1968.

Schmid H., « Probabilistic part-of-speech tagging using decision trees », *Proceedings of International Conference on New Methods in Language Processing*, vol. 12, Manchester, UK, 1994.

Schwarm S., Ostendorf M., « 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, 2005.

Si L., Callan J., « A statistical model for scientific readability », *Proceedings of the Tenth International Conference on Information and Knowledge Management*, ACM New York, NY, USA, p. 574-576, 2001.

Smith E., « Devereaux readability index », *The Journal of Educational Research*, vol. 54, n° 8, p. 289-303, 1961.

Stenner A., « Measuring reading comprehension with the lexile framework », *Fourth North American Conference on Adolescent/Adult Literacy*, 1996.

Taylor W., « Cloze procedure : A new tool for measuring readability », *Journalism quarterly*, vol. 30, n° 4, p. 415-433, 1953.

Tharp J., « The Measurement of Vocabulary Difficulty », *The Modern Language Journal*, vol. 24, n° 3, p. 169-178, 1939.

Tufféry S., *Data mining et statistique décisionnelle l'intelligence des données*, Éd. Technip, Paris, 2007.

Uitdenbogerd S., « Readability of French as a foreign language and its uses », *Proceedings of the Australian Document Computing Symposium*, p. 19-25, 2005.

van Oosten P., Hoste V., Tanghe D., « A Posteriori Agreement as a Quality Measure for Readability Prediction Systems », *in* A. Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*, vol. 6609 of *Lecture Notes in Computer Science*, Springer, Berlin / Heidelberg, p. 424-435, 2011.

Vitu F., O'Regan J., Mittau M., « Optimal landing position in reading isolated words and continuous text », *Perception & Psychophysics*, vol. 47, n° 6, p. 583-600, 1990.

Vogel M., Washburne C., « An objective method of determining grade placement of children's reading material », *The Elementary School Journal*, vol. 28, n° 5, p. 373-381, 1928.

Witten I., Frank E., *Data Mining : Practical machine learning tools and techniques*, Morgan Kaufmann, San Francisco, 2005.