RECITAL 2009, Senlis, 24–26 juin 2009

Modèles statistiques pour l'estimation automatique de la difficulté de textes de FLE

Thomas François
Aspirant FNRS
Centre de Traitement Automatique du Langage
Université Catholique de Louvain
thomas.francois@uclouvain.be

Résumé. La lecture constitue l'une des tâches essentielles dans l'apprentissage d'une langue étrangère. Toutefois, la découverte d'un texte portant sur un sujet précis et qui soit adapté au niveau de chaque apprenant est consommatrice de temps et pourrait être automatisée. Des expériences montrent que, pour l'anglais, l'utilisation de classifieurs statistiques permet d'estimer automatiquement la difficulté d'un texte. Dans cet article, nous proposons une méthodologie originale comparant, pour le français langue étrangère (FLE), diverses techniques de classification (la régression logistique, le bagging et le boosting) sur deux corpus d'entraînement. Il ressort de cette analyse comparative une légère supériorité de la régression logistique multinomiale.

Abstract. Reading is known to be an essential task in language learning, but finding the appropriate text for every learner is far from easy. In this context, automatic procedures can support the teacher's work. Some works on English reveal that it is possible to assess the readability of texts using statistical classifiers. In this paper, we present an original approach comparing various classification techniques, namely logistic regression, bagging and boosting on two training corpora. The results show a slight superiority for multinomial logistic regression over bagging or boosting.

Mots-clés : lisibilité, régression logistique, bagging, boosting, modèle de langue.

Keywords: readability, logistic regression, bagging, boosting, language model.

1 Introduction

Aujourd'hui, sous l'effet de l'élargissement européen et d'un accroissement de la mobilité, le secteur de l'enseignement des langues se trouve en pleine croissance et l'offre peine à suivre la demande. Le développement de l'ALAO (Apprentissage des Langues Assisté par Ordinateur) vise à répondre à ces nouveaux besoins, notamment en automatisant certaines tâches répétitives inhérentes à l'enseignement des langues. Parmi celles-ci, la recherche de documents authentiques portant sur un sujet précis et adapté au niveau des apprenants constitue une tâche coûteuse en temps. C'est pourquoi nous proposons dans cet article un modèle de lisibilité apte à estimer rapidement la difficulté d'un texte sur la base d'algorithmes de classification. Il a pour objectif de faciliter aussi bien le travail des concepteurs de supports pédagogiques que celui des professeurs de FLE amenés à utiliser l'Internet pour la préparation de leur cours.

Nous discutons des recherches antérieures en lisibilité dans la section 2, avant de présenter la méthodologie propre à notre approche. Celle-ci se base sur un corpus et une échelle de difficulté, présentés dans la section 3. De chacun des textes de ce corpus est ensuite extrait une série de variables linguistiques décrites à la section 4 et qui sont ensuite utilisées au sein de modèles statistiques prédictifs dont les détails techniques sont résumés à la section 5. La section 6 détaille quelques remarques propres à notre implémentation et la section 7 discute les résultats obtenus. Enfin, nous concluons avec la section 8 en avançant diverses pistes de recherche.

2 Recherches en lisibilité

La première méthode visant à évaluer la difficulté d'un texte pour un lectorat déterminé fut le jugement d'expert, qui recoure à des critères non explicites. Toutefois, dans les années 20, les premières méthodes reproductibles apparurent avec les travaux de Lively and Pressey (1923), à l'origine de la première formule de lisibilité. Par la suite, le domaine s'est développé et a produit, pour l'anglais, une série de formules basées sur des indices lexicaux et syntaxiques (Flesch, 1948; Chall & Dale, 1995). Par contre, il faut attendre 1956 et l'ouvrage d'André Conquet, *La lisibilité* (1971) pour que le monde francophone découvre ce champ de recherche. Les premières formules pour le français sont dues à Kandel et Moles (1958) et de de Landsheere (1963), mais elles ne constituent encore qu'une adaptation de la formule de Flesch, sans que soit pris en compte l'ensemble des spécificités de la langue française.

La première formule spécifique au français est due à Henry (1975). Ce chercheur expérimente un grand nombre de variables linguistiques dont il tire trois formules, de loin les plus utilisées et plus les abouties pour le français. Étrangement, bien peu de travaux suivent ces premières percées. On ne compte que Richaudeau (1979), qui préfère substituer aux formules de lisibilité un critère d'efficacité linguistique développé à partir d'expériences sur la mémoire à court terme, et Mesnager (1989), qui conçoit la formule la plus récente pour le français avec comme population cible, les enfants. Pour Bossé-Andrieu (1993), ce manque d'intérêt s'explique par des raisons culturelles : l'idée de mesurer un texte par des moyens objectifs constituerait une approche trop pragmatique pour l'esprit français.

Quoi qu'il en soit, il faut noter que si peu de travaux en lisibilité ont porté sur le français langue première (L1), on en compte encore moins qui se sont attachés aux particularités du FLE. Cornaire (1988) a testé la validité de la formule de Henry pour le FLE et, plus récemment, Uitdenbogerd (2005) s'est intéressée, au travers des cognates, à la prise en compte de la proximité lexicale entre deux langues et a développé une formule de FLE destinée à des apprenants anglophones.

Confrontés à ce relatif oubli du domaine, nous nous sommes tournés vers le monde anglo-saxon où la lisibilité a connu un renouveau récent sous l'impulsion du TAL et de techniques d'apprentissage automatique. Il est désormais possible de tester la capacité prédictive de variables plus complexes. Ainsi, Collins-Thompson *et al.* (2005) ont proposé un classifieur bayésien naïf qui correspond à un modèle unigramme lissé, montrant par là qu'il était possible de remplacer les listes de mots les plus communs par des modèles de langue. De leur côté, Schwarm et Ostendorf (2005) ont utilisé des support vector machines (SVM) afin de combiner certaines des variables classiques en lisibilité avec, d'une part, une série de modèles de langue trigrammes (un modèle par niveau de difficulté) et, d'autre part, des caractéristiques syntaxiques basées sur des arbres de dérivation. Heilman *et al.* (2007) ont enrichi le modèle unigramme de en lui ajoutant

la reconnaissance de structures syntaxiques, en vue d'estimer la difficulté de textes en anglais comme langue étrangère. Par la suite, ils ont amélioré la capacité prédictive de leurs diverses variables en utilisant des méthodes de régression (Heilman *et al.*, 2008). C'est sur la base de ces différents travaux que nous avons étudié les possibilités d'adaptation de diverses techniques à la lisibilité du FLE.

3 Description du corpus

La première étape dans le développement d'une nouvelle formule de lisibilité consiste à collecter un corpus de textes qui soient déjà catégorisés selon une échelle de difficulté. Dans un contexte européen, il nous est apparu logique d'opter pour l'échelle qui sert de référence pour l'ensemble des programmes d'éducation en langue étrangère, à savoir le *Cadre européen commun de référence pour les langues* (CECR) (Conseil de l'Europe, 2001) . Le CECR a défini six niveaux : A1 (le plus bas), A2, B1, B2, C1 et C2 (le plus élevé)

Pour rassembler en nombre suffisant des textes étiquetés selon la même échelle, nous avons utilisé des manuels de FLE qui, depuis la création du CECR, ont connu une certaine standardisation en termes de difficulté. Il est dès lors possible de constituer un corpus de documents étiquetés par des experts en FLE. Cependant, seuls certains manuels sont adaptés aux objectifs de notre recherche et tout leur contenu n'est pas utilisable. C'est pourquoi nous avons défini des critères de sélection suivants :

– Les manuels doivent être postérieurs à 2001, date de la publication du CECR. Cette restriction permet également de s'assurer que le langage contenu dans les textes soit proche du français actuel.
– Seuls des manuels destinés à des adultes ou des jeunes gens sont retenus puisqu'il s'agit de la population de lecteurs visée par notre formule.
– Parmi les textes, n'ont été conservés que ceux qui sont constitués de phrases complètes et qui dépendent d'une tâche de compréhension à la lecture.

En respectant ces critères, nous avons rassemblé près de 2 000 textes, représentant plus de 500 000 mots, et qui couvrent des sujets divers : extraits de littérature, articles de journaux, dialogues, recettes de cuisine... L'objectif est de permettre à la formule obtenue une capacité de généralisation optimale tout en repérant les types de textes pour lesquels elle ne s'applique pas.

4 Variables lexicales et syntaxiques

Les travaux en lisibilité ont toujours visé à paramétrer les textes sous la forme de variables qui constituent de bons indices de la difficulté (c.-à-d. qui soient fortement corrélées avec cette difficulté). Les approches classiques, qui n'ont recouru qu'à des variables de types lexical et syntaxique, furent largement critiquées par des cognitivistes tels que Kintsch and Vipond (1979) et Kemper (1983). Ces auteurs soulignèrent l'importance de prendre en compte les aspects conceptuels des textes, tels que les relations entre phrases ou la charge inférentielle. Bien que théoriquement fondées, ces approches ne menèrent pas à des modèles aisément reproductibles et automatisables, ce qui explique le retour à des variables lexicales et syntaxiques, en profitant toutefois des avancées du TAL. Notre recherche étant encore loin de son terme, nous n'avons encore expérimenté que quelques variables, décrites ci-dessous.

4.1 Les variables classiques

Parmi les prédicteurs couramment utilisés, nous avons testé le rapport type-token, le nombre moyen de lettres par mot et la longueur moyenne des phrases. Nous n'avons conservé que les deux derniers, qui présentaient une corrélation élevée avec la difficulté. De plus, nous avons repris chez Henry (1975, p. 85) cinq variables de dialogue : le rapport des pronoms personnels de dialogue (1re et 2e personnes) aux pronoms (PPD), la proportion d'interjections (PI), le pourcentage de points d'exclamation et de points d'interrogation par rapport au nombre total de signes de ponctuation (PPEI) et la présence de guillemets (BINGUI).

4.2 La fréquence lexicale mesurée à l'aide d'un modèle de langue

La fréquence des mots d'un texte est considérée depuis longtemps comme un excellent indice de la complexité lexicale (Howes & Solomon, 1951). Or, comme l'ont montré les travaux de Collins-Thompson *et al.* (2005), les modèles de langue peuvent remplacer avantageusement les listes de vocabulaire comme mode de paramétrisation. C'est pourquoi nous avons considéré la probabilité d'un texte T (avec N mots w_i) comme un indice de la complexité lexicale, calculé sur base de l'équation 1 :

$$P(T) = P(w_1)P(w_2 \mid w_1) \cdots P(w_n \mid w_1, w_2, \ldots, w_{n-1}) \tag{1}$$

Cette équation soulève deux problèmes :

1. La difficulté, bien connue, d'estimer correctement l'ensemble de ces probabilités conditionnelles. Nous avons opté pour un modèle par unigramme lissé puisque, d'après Collins-Thompson *et al.* (2005), il semble donner de bons résultats. La probabilité d'un texte est donc réduite à l'équation suivante :

   $$P(T) = exp(\sum_{i=1}^{n} \log[p(w_i)]) \tag{2}$$

   Le résultat doit ensuite être normalisé en divisant par le nombre N de mots, afin de le rendre indépendant de la longueur du texte. Davantage de détails concernant l'origine et le lissage des probabilités sont décrits dans la section 6.

2. L'unité linguistique à prendre en compte. Celle traditionnellement utilisée en lisibilité est la forme fléchie, mais la nature flexionnelle du français laisse supposer que le lemme pourrait constituer une meilleure alternative. Par ailleurs, d'un point de vue théorique, l'emploi des formes fléchies sous-entend que le lecteur n'est pas capable d'inférer le sens d'une forme inconnue même s'il connaît une autre forme de ce même mot, une position qui paraît critiquable pour la majorité des formes régulières.

   Dans le but d'éclaircir cette question à l'aide de résultats concrets, nous avons entraîné trois modèles de langue : un premier basé sur des lemmes[1] (LM1), un second considérant les formes fléchies désambiguïsées en fonction de leur catégorie(LM2) et un dernier recourant simplement aux formes fléchies (LM3)[1]. Cependant, l'expérience n'a pas été très informative, puisque ces trois variables sont corrélées de manière assez similaire avec la difficulté dans nos différents sous-corpus de test. C'est pourquoi nous avons conservé les trois variables lors de cette première étape.

[1] Pour ces trois modèles, les textes ont été analysés à l'aide du TreeTagger (Schmid, 1994).

4.3 Mesure de la complexité verbale

Une caractéristique intéressante de l'enseignement des langues étrangères est la séquentialité de l'apprentissage, c'est-à-dire qu'on apprend certaines formes linguistiques avant d'autres. C'est d'autant plus vrai pour la conjugaison où certains temps ou modes sont systématiquement étudiés avant d'autres. Par conséquent, il est possible d'utiliser cette information comme prédicteur de la difficulté de textes de FLE, étant donné qu'un apprenant débutant est peu susceptible de connaître le temps, le mode et l'aspect d'une forme verbale au passé simple et risque donc de mal en saisir le sens.

Nous avons ainsi défini 11 variables binaires comme indices de cette complexité verbale : le conditionnel, le futur, l'impératif, l'indicatif imparfait, l'infinitif, le participe passé, le participe présent, le passé simple, l'indicatif présent, le subjonctif présent et le subjonctif imparfait.

5 Présentation des modèles statistiques

À l'issue de cette étape de paramétrisation, nous avons obtenu, pour chacun des textes du corpus, 20 prédicteurs associés à une classe qui représente la difficulté. Les raisons qui justifient de considérer la difficulté comme une variable catégorielle sont discutées dans François (2009) et se résument à un manque d'adéquation aux données des modèles basés sur une variable dépendante continue ou ordinale. En effet, nous avons montré que la régression logistique multinomiale se révélait supérieure à des classifieurs basés sur la régression linéaire, la régression logistique ordinale ou les arbres de décision (François, 2009).

Par conséquent, nous avons voulu comparer la régression logistique multinomiale à deux autres techniques de classification, considérées comme parmi les plus efficaces : le bagging et le boosting. Dans la suite de cette section, nous présentons brièvement ces trois techniques statistiques.

5.1 La régression logistique multinomiale

La régression logistique a d'abord été développée pour des données binaires et, comme toutes les techniques de régression, elle vise à modéliser l'espérance conditionnelle $E(Y \mid X)$, c.-à-d. la valeur moyenne attendue pour Y étant donné une valeur de X. À la différence de la régression linéaire, qui modélise cette espérance à l'aide d'une droite, la régression logistique recourt à une courbe en S (sigmoïde), évitant ainsi d'attribuer à Y des valeurs sortant de l'intervalle $[0, 1]$.

La régression multinomiale (RLM) constitue une généralisation de cette technique pour un problème à K classes et se réalise à l'aide d'un modèle constitué de $K-1$ sigmoïdes. Chacune d'elles compare la classe k à une catégorie de référence (souvent la première), afin de retomber sur le cas binaire. Ainsi, pour chacune de ces paires de classes (Y_j, Y_1), il existe une fonction décrite sous forme matricielle par l'équation suivante (Agresti, 2002, p. 268) :

$$log\frac{P(Y=k \mid X=\mathbf{x})}{P(Y=1 \mid X=\mathbf{x})} = \alpha_k + \boldsymbol{\beta}_k^T\mathbf{x} \qquad k=2,\cdots,K \tag{3}$$

où $\mathbf{x}$ est un vecteur observation, α_k et $\boldsymbol{\beta}_k^T$ représentent les paramètres du modèle pour la classe k. Sur la base de ces $K-1$ équations, il est possible de calculer la probabilité qu'un texte

appartienne au niveau de difficulté k pour un vecteur $\mathbf{x}$ donné. Cette probabilité est donnée par l'équation suivante[2] (Agresti, 2002, p. 271) :

$$P(Y = k \mid X = \mathbf{x}) = \frac{\exp(\alpha_k + \boldsymbol{\beta}_k^T \mathbf{x})}{1 + \sum_{k=2}^{K} \exp(\alpha_k + \boldsymbol{\beta}_k^T \mathbf{x})} \tag{4}$$

L'estimation des paramètres d'un tel modèle s'effectue par maximum de vraisemblance à l'aide d'une procédure décrite dans Agresti (2002, p. 192).

5.2 Le bagging et le boosting

Le bagging et le boosting partagent le même postulat théorique : mieux vaut un ensemble de classifieurs qu'un seul. La difficulté de cette approche provient de la nécessité de disposer d'un corpus différent pour entraîner chacun des classifieurs. C'est dans leur manière de répondre à cette problématique que les deux approches se distinguent.

Le bagging, une technique développée par Breiman (1996), va générer aléatoirement N échantillons par rééchantillonnage, qui permettront d'entraîner N classifieurs. Par la suite, le niveau d'un texte peut être défini comme la catégorie majoritaire lors du vote de ces N classifieurs. Le bagging présente l'avantage de réduire la variance d'un modèle et s'applique donc particulièrement bien à des classifieurs instables, c.-à-d. qui peuvent fortement varier en fonction des données d'entraînement, tels que les arbres de décision.

Le boosting, mis au point par Freund and Schapire (1997), présente la particularité d'être adaptatif, car il se concentre sur les données difficiles à modéliser. C'est pourquoi l'algorithme le plus connu s'appelle AdaBoost (Adaptative Boosting) et fonctionne de la manière suivante :

- Un poids $w_i = \frac{1}{N}$ est attribué à chacune des N observations du corpus d'entraînement. Sur la base de ce corpus, un classifieur $h_t(\mathbf{x})$ est entraîné et une estimation de son taux d'erreur ϵ_t est obtenue (pour le détail des formules, se reporter à Meir and Ratsch (2003)).
- Sur la base de cette estimation de l'erreur, on peut calculer le poids α_t du classifieur, avant de réévaluer le poids w_i de chacune des observations, donnant plus d'importance aux données mal classées et moins à celles qui ont été correctement catégorisées.
- Ces deux étapes sont itérées T fois, après quoi on obtient un ensemble de T classifieurs qui attribuent à un texte un niveau de difficulté par un vote pondéré, ce qui revient à l'équation suivante (Meir & Ratsch, 2003, p. 118):

$$f_{Ens}(\mathbf{x}) = \sum_{t=1}^{T} \alpha_t h_t(\mathbf{x}) \tag{5}$$

Le boosting ne convient normalement qu'à une variable dépendante de type binaire. Nous avons adapté cette technique pour un problème à K classes en recourant à la stratégie du un contre tous. Celle-ci développe K ensembles, en opposant chaque classe k au reste des données. On obtient donc K prédicteurs qui, pour un texte donné, vont chacun estimer la probabilité que ce texte appartienne à la classe k. Reste alors à attribuer à ce texte la classe pour laquelle $f_{Ens,k}(\mathbf{x})$ est la plus élevée.

[2] Notons que pour la catégorie de référence ($k = 1$), α_1 et $\boldsymbol{\beta}_1^T = 0$. Aussi, lors du calcul de la probabilité qu'un texte appartienne à cette catégorie de référence, le numérateur vaut $\exp(0) = 1$.

6 Implémentations

Après avoir décrit les aspects théoriques de notre modèle, nous présentons quelques détails de son implémentation, envisageant d'abord le modèle de langue, avant de s'arrêter sur la sélection des variables linguistiques.

Notre modèle de langue a nécessité de disposer d'une liste des lemmes et formes fléchies du français auxquels est associée leur fréquence d'apparition dans la langue. Nous avons utilisé *Lexique3*, un lexique développé par New et al. (2001) qui contient plus de 50 000 lemmes dont les fréquences ont été estimées sur un corpus de sous-titres de films rassemblant plus de 50 millions de mots. À partir de ces fréquences, nous avons estimé des probabilités par maximum de vraisemblance avant de les lisser à l'aide de l'algorithme *Simple Good-Turing* décrit par Gale et Sampson (1995). Cette étape est nécessaire pour être capable d'attribuer une probabilité aux mots n'ayant pas été observés dans le corpus d'apprentissage.

En ce qui concerne la sélection des variables, il convenait de déterminer, parmi les 20 prédicteurs obtenus à l'issue de la phase de paramétrisation, lesquels représentent les meilleurs prédicteurs de la complexité d'un texte. Pour le bagging et le boosting, qui sont basés sur des arbres de décision, la sélection de variables s'opère automatiquement à chaque noeud de chaque arbre grâce au critère de Gini. Par contre, pour la régression logistique multinomiale, il est nécessaire d'obtenir le modèle le plus parcimonieux possible, étant donné le nombre élevé de paramètres[3]. Pour ce faire, nous avons utilisé un algorithme de sélection pas à pas des variables, qui compare les différents modèles possibles et sélectionne celui qui est à la fois efficace et parcimonieux. Ce modèle est celui qui obtient la valeur AIC (Akaike's Information Criterion) la plus élevée, celle-ci étant définie comme équivalent à $-2 * \text{log-vraisemblance} + 2p$, où p correspond au nombre de paramètres du modèle et la log-vraisemblance s'obtient à l'aide d'un calcul détaillé dans Hosmer and Lemeshow (1989).

7 Résultats

Une fois cette méthodologie mise en place, deux sous-corpus, obtenus par un rééchantillonnage du corpus global, ont été constitués. Le premier d'entre eux comprend 288 textes répartis selon les 6 niveaux du CECR, alors que le second rassemble 437 textes étalés sur les 9 niveaux de difficulté[4] suivants : A1, A1+, A2, A2+, B1, B1+, B2, C1 et C2. Chaque niveau comprenait 50 textes, formant nos deux corpus desquels ont été exclus respectivement 12 et 13 données aberrantes, définis comme des données situées au-delà de 3 écarts-types de leur moyenne (ce qui correspond à un α de 0,0026).

Ensuite, pour chacun des deux corpus, trois classifieurs ont été entraînés, respectivement par RLM, bagging et boosting. Les modèles obtenus ont été évalués à l'aide des trois mesures suivantes : la corrélation (r de Pearson) entre les niveaux réels des textes et les prédictions, l'exactitude des prédictions (définie comme le rapport du nombre de textes correctement classés sur le nombre total de textes) et l'exactitude contiguë[5].

[3]Ce nombre dépend du nombre K de classes et du nombre X de variables indépendantes selon la relation suivante : nombre de paramètres $= (K-1)(X+1)$.

[4]L'objectif de cette seconde échelle est de modéliser plus finement les premières étapes de l'apprentissage où les différences se font davantage sentir au sein d'un même niveau.

[5]Cette mesure est définie par Heilman *et al.* (2008) comme la proportion de prédictions qui s'éloigne au plus

Mesure	RLM	bagging	boosting	Mesure	RLM	bagging	boosting
Résultats sur les échantillons d'apprentissage							
Sous-corpus à 6 niveaux de difficulté				**Sous-corpus à 9 niveaux de difficulté**			
Corrélation	$0,70$	1	$0,97$	Corrélation	$0,74$	1	$0,99$
Exactitude	50%	100%	97%	Exactitude	41%	100%	97%
Exac. contiguë	76%	100%	98%	Exac. contiguë	66%	100%	98%
Résultats sur les échantillons de test							
Sous-corpus à 6 niveaux de difficulté				**Sous-corpus à 9 niveaux de difficulté**			
Corrélation	$0,62$	$0,60$	$0,64$	Corrélation	$0,72$	$0,69$	$0,68$
Exactitude	41%	37%	40%	Exactitude	32%	29%	29%
Exac. contiguë	71%	70%	70%	Exac. contiguë	63%	65%	64%

TAB. 1 – Estimation, par une procédure de validation croisée, du coefficient de Pearson's, de l'exactitude et de l'exactitude contiguë pour les différents classifieurs.

Pour chacun des classifieurs, nous avons effectué une procédure de validation croisée à dix échantillons afin d'estimer plus précisément les trois mesures d'évaluation. La Table 1 présente les résultats obtenus sur nos deux sous-corpus. Notons que, sur les deux corpus et pour les différentes mesures d'évaluation, les trois classifieurs présentent des résultats qui ne sont pas significativement différents. Il est donc impossible de conclure à la supériorité de l'un des modèles statistiques, même si la méthode qui semble se comporter le mieux est la RLM. Sachant de plus que le temps nécessaire à l'entraînement d'un tel modèle est considérablement plus réduit que celui pour le bagging et surtout le boosting, il nous semble que la RLM constitue le meilleur choix aussi bien au niveau de l'optimisation du temps de calcul qu'au niveau de l'efficacité du modèle. Cette conclusion concorde avec nos résultats précédents (François, 2009).

En termes d'efficacité, l'exactitude obtenue (respectivement 41% et 32% pour la RLM) paraît peu satisfaisante. L'exactitude contiguë révèle cependant que les prédictions se concentrent assez bien autour de la diagonale de la matrice de confusion, ce qui signifie que les textes sont plus ou moins bien classés à un niveau près. D'ailleurs, si nous comparons ces résultats à la seule approche similaire pour le français, réalisée par Collins-Thompson et Callan (2005) (qui obtiennent une corrélation de 0,64 entre leur variable dépendante à 5 classes et leurs prédicteurs), nous nous situons dans le même ordre de grandeur, avec un r de 0,62 pour 6 classes et de 0,72 pour 9 classes.

Cette similarité dans les résultats confirme que prédire automatiquement la difficulté d'un texte constitue une tâche ardue. Les meilleurs taux pour l'anglais, où la recherche dans ce domaine est plus avancée, ne dépassent d'ailleurs pas 52% d'exactitude contiguë pour 12 classes (Heilman *et al.*, 2008). Le problème s'explique aussi par le fait que les psychologues n'aient pas encore réussi à s'accorder sur une description explicite des facteurs qui font la complexité d'un texte. Dès lors, les chercheurs en lisibilité en sont réduits à vérifier expérimentalement l'importance des différentes variables.

d'un niveau par rapport à celui qui a été attribué au texte par un humain. Son emploi se justifie par la difficulté qu'ont divers experts humains à accorder leur jugement sur un ensemble de textes. Il ne faut toutefois pas oublier qu'elle fournit des valeurs exagérément optimistes lorsqu'il y a peu de catégories.

8 Perspectives de recherche

En vue d'une amélioration des performances, les deux axes suivants seront poursuivis. D'une part, il s'agira d'expérimenter un plus grand nombre d'indices de difficulté. Nous pensons ainsi recourir à un modèle lexical par n-grammes, distinguer entre mots concrets et abstraits ou encore utiliser diverses informations obtenues sur la base d'un analyseur syntaxique, telles que le nombre moyen de nœuds par phrase, le nombre moyen de propositions par phrase ou encore la profondeur lexicale moyenne (Bormuth, 1966).

D'autre part, une grande part de la perte de performance s'explique par le bruit contenu dans le corpus. En effet, les matériaux récoltés au sein de manuels de FLE ont le défaut de présenter une grande variabilité au sein d'un même niveau. Il convient donc de développer un moyen de réduire ce bruit. Pour l'instant, nous envisageons d'entraîner les modèles en plusieurs passes au cours desquelles les textes trop mal prédits seraient exclus du corpus d'apprentissage.

Par ailleurs, comme nos expérimentations montrent que l'emploi de diverses méthodes de classification ne débouche pas sur une amélioration sensible des résultats, nous envisageons de limiter les investigations dans ce sens à une seule autre méthode de classification bien connue, les SVM. Ces différentes expériences permettront d'obtenir une estimation précise de l'importance de chacun de ces trois facteurs que sont le corpus d'apprentissage, les prédicteurs et les techniques de classification dans la construction de nouvelles formules de lisibilité basées sur des techniques de TAL.

Références

AGRESTI A. (2002). *Categorical Data Analysis. 2nd edition*. New York: Wiley-Interscience.

BORMUTH J. (1966). Readability: A new approach. *Reading research quarterly*, p. 79–132.

BOSSÉ-ANDRIEU J. (1993). La question de la lisibilité dans les pays anglophones et les pays francophones. *Technostyle, Association canadienne des professeurs de rédaction technique et scientifique*, **11**(2), 73–85.

BREIMAN L. (1996). Bagging predictors. *Machine learning*, **24**(2), 123–140.

CHALL J. & DALE E. (1995). *Readability Revisited: The New Dale-Chall Readability Formula*. Cambridge: Brookline Books.

COLLINS-THOMPSON K. & CALLAN J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, **56**(13), 1448–1462.

CONQUET A. (1971). *La lisibilité*. Paris: Assemblée Permanente des CCI de Paris.

CONSEIL DE L'EUROPE . (2001). *Cadre européen commun de référence pour les langues*.

CORNAIRE C. (1988). La lisibilité : essai d'application de la formule courte d'Henry au français langue étrangère. *Canadian Modern Language Review*, **44**(2), 261–273.

DE LANDSHEERE G. (1963). Pour une application des tests de lisibilité de Flesch à la langue française. *Le Travail Humain*, **26**, 141–154.

FLESCH R. (1948). A new readability yardstick. *Journal of Applied Psychology*, **32**(3), 221–233.

FRANÇOIS T. (2009). Combining a Statistical Language Model with Logistic Regression to Predict the Lexical and Syntactic Difficulty of Texts for FFL. In *Proceedings of the EACL 2009 Student Research Workshop*, p. 19–27, Athens, Greece.

GALE W. & SAMPSON G. (1995). Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, **2**(3), 217–237.

HEILMAN M., COLLINS-THOMPSON K., CALLAN J. & ESKENAZI M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of NAACL HLT*, p. 460–467.

HEILMAN M., COLLINS-THOMPSON K. & ESKENAZI M. (2008). An analysis of statistical models and features for reading difficulty prediction. *Association for Computational Linguistics*, **The 3rd Workshop on Innovative Use of NLP for Building Educational Applications**, 1–8.

HENRY G. (1975). *Comment mesurer la lisibilité*. Labor.

HOSMER D. & LEMESHOW S. (1989). *Applied Logistic Regression*. New York: Wiley.

HOWES D. & SOLOMON R. (1951). Visual duration threshold as a function of word probability. *Journal of Experimental Psychology*, **41**(40), 1–4.

KANDEL L. & MOLES A. (1958). Application de l'indice de Flesch à la langue française. *Cahiers Études de Radio-Télévision*, **19**, 253–274.

KEMPER S. (1983). Measuring the inference load of a text. *Journal of Educational Psychology*, **75**(3), 391–401.

KINTSCH W. & VIPOND D. (1979). Reading comprehension and readability in educational practice and psychological theory. *Perspectives on Memory Research*, p. 329–366.

LIVELY B. & PRESSEY S. (1923). A method for measuring the "vocabulary burden" of textbooks. *Educational Administration and Supervision*, **9**, 389–398.

MEIR R. & RATSCH G. (2003). An introduction to boosting and leveraging. *Lecture Notes in Computer Science*, **2600**, 118–183.

MESNAGER J. (1989). Lisibilité des textes pour enfants: un nouvel outil ? *Communication et Langages*, **79**, 18–38.

NEW B., PALLIER C., FERRAND L. & MATOS R. (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE. *L'Année Psychologique*, **101**, 447–462.

RICHAUDEAU F. (1979). Une nouvelle formule de lisibilité. *Communication et Langages*, **44**, 5–26.

SCHAPIRE R. & FREUND Y. (1997). A decision theoretic generalization of on-line learning and an application to boosting. *Journal Computer and System Sciences*, **55**, 119–139.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, volume 12: Manchester, UK.

SCHWARM S. & OSTENDORF M. (2005). Reading level assessment using support vector machines and statistical language models. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, p. 523–530.

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