Classification automatique de dictées selon leur niveau de difficulté de compréhension et orthographique

Adeline Müller[1] Thomas François[1, 2] Sophie Roekhaut[1] Cédrick Fairon[1]
(1) CENTAL, IL&C, UCL, 1348 Louvain-la-Neuve, Belgium
(2) Chargé de recherche FNRS
adeline.muller@student.uclouvain.be, thomas.francois@uclouvain.be,
sroekhaut@altissia.com, cedrick.fairon@uclouvain.be

RÉSUMÉ

Cet article présente une approche visant à évaluer automatiquement la difficulté de dictées en vue de les intégrer dans une plateforme d'apprentissage de l'orthographe. La particularité de l'exercice de la dictée est de devoir percevoir du code oral et de le retranscrire via le code écrit. Nous envisageons ce double niveau de difficulté à l'aide de 375 variables mesurant la difficulté de compréhension d'un texte ainsi que les phénomènes orthographiques et grammaticaux complexes qu'il contient. Un sous-ensemble optimal de ces variables est combiné à l'aide d'un modèle par machines à vecteurs de support (SVM) qui classe correctement 56% des textes. Les variables lexicales basées sur la liste orthographique de Catach (1984) se révèlent les plus informatives pour le modèle.

ABSTRACT

Automatic classification of dictations according to their complexity for comprehension and writing production.

This paper introduces a new approach that aims to automatically assess the difficulty of texts intended to be dictation exercises within a web learning plateform. The most remarkable feature about the dictation exercise is the written transcription of a perceived oral code. To model this process, we take into account 375 features aiming at measuring the difficulty of a text at the comprehension level as well as at the spelling and grammatical level. Based on an optimal subset of features, three support vector machine (SVM) models are trained, the best of which is able to correctly classify 56% of the dictations. The most predictive features are those based on the Catach (1984)'s word list.

MOTS-CLÉS : dictée, lisibilité, orthographe, ALAO.

KEYWORDS: dictation, readability, spelling, CALL.

1 Introduction

Malgré les nombreux débats dont elle a fait l'objet, la dictée reste une activité incontournable à l'école (Delabarre & Devillers, 2015). Les manuels et les programmes proposent dès lors différents types de dictées, essayant de mettre en valeur cette activité comme un apprentissage permettant aux élèves d'être acteurs et non seulement « copistes » (Brissaud & Mortamet, 2015). Toutefois, la dictée comme activité de drill reste nécessaire au renforcement des apprentissages et gagnerait à être automatisée afin de libérer du temps d'enseignement pour des activités plus réflexives. C'est

un objectif qui entre dans le cadre de l'enseignement des langues assisté par ordinateur (ELAO) et du traitement automatique du langage (TAL) et qui a déjà été abordé à plusieurs reprises au sein de ces domaines (Santiago Oriola, 1998; Ruggia, 2000; Beaufort & Roekhaut, 2011). Parmi ces travaux, les efforts ont surtout porté sur l'implémentation d'une méthode précise et pédagogique de correction automatique des dictées. Notre article s'attaque à une autre facette de l'automatisation des dictées, à savoir la sélection des dictées. Si les enseignants sont généralement capables de sélectionner un texte à dicter adapté à leur public, une plateforme d'apprentissage automatisée requiert, de son côté, de disposer d'un corpus de textes, préalablement annoté selon le niveau de difficulté par des experts. Cette caractéristique limite l'adaptabilité de ce type de système, qui, par exemple, ne pourra sélectionner des textes d'actualité ou adaptés au profil d'un apprenant en termes de connaissances grammaticales ou thématique que si sa base de dictées en contient. C'est pourquoi cet article explore les possibilités de déterminer automatiquement le niveau de difficulté d'une dictée.

L'article est articulé en trois sections principales. Tout d'abord, nous décrirons les études liées à l'évaluation de textes et aux difficultés orthographiques (Section 2). Ensuite, nous présenterons la méthodologie utilisée pour élaborer notre modèle, en détaillant le corpus utilisé, les variables mises au point pour capturer les phénomènes orthographiques ardus, et le modèle statistique qui sert à les combiner (Section 3). Enfin, les résultats obtenus seront exposés et discutés à la section 4.

2 Contexte

Réaliser une dictée constitue une tâche particulièrement complexe, parce qu'elle nécessite de changer de canal de communication. Le texte, oralisé par un enseignant ou un agent informatique, doit être compris et transposé par écrit par l'étudiant. Comme l'ont montré Schelstraete & Maillart (2004), ce passage de l'oral vers l'écrit constitue une activité cognitive complexe, qui peut entrainer, dans certains cas, une surcharge cognitive, en raison de trois grands facteurs :

— Le ralentissement du débit lors du passage à l'écrit : cela suppose que le scripteur doit garder en mémoire différentes informations assez longtemps.
— Le contrôle de l'activité grapho-motrice : la réalisation graphique elle-même peut demander de nombreuses ressources attentionnelles, en particulier chez les scripteurs débutants.
— L'orthographe : gérer les problèmes d'orthographe demande de faire appel à des ressources cognitives, même chez les scripteurs «experts».

Selon les expériences de Bourdin (1994), les deux derniers facteurs sont les plus problématiques, même pour les scripteurs adultes. Si l'évaluation de l'activité grapho-motrice sort du champ de notre étude, notre modèle vise par contre à prendre en considération les deux autres aspects.

Le premier d'entre eux peut être associé à la capacité de compréhension d'un texte et à sa représentation en mémoire. La discipline qui vise à évaluer automatiquement la difficulté de compréhension d'un texte est la lisibilité. De nombreux travaux ont porté sur l'évaluation automatique de la difficulté des textes à la lecture. Pour l'anglais, on compte des formules bien connues telles que celles de Flesch (1948), de Dale & Chall (1948) ou, plus récemment, de Collins-Thompson & Callan (2005), Feng *et al.* (2010) ou Vajjala & Meurers (2012). Tandis que les premières se basent sur des variables simples (nombre de mots, etc.), les secondes s'appuient sur des techniques de TAL et combinent un nombre plus important de variables de différents niveaux linguistiques (lexical, syntaxique, sémantique ou discursif). Pour le français, les formules de Kandel & Moles (1958) et Henry (1975) sont plutôt

typiques du courant classique, tandis que des modèles reposant sur le TAL ont été développés par François & Fairon (2013) ou Dascalu (2014). Bien que la lisibilité prenne aussi en compte la phase de décodage du code graphique, absent dans notre cas, nous pensons que ce type de variables pourrait être transposé pour évaluer le niveau de compréhensibilité de nos dictées, d'autant que, à notre connaissance, il n'existe pas de recherches spécifiquement dédiées à l'évaluation automatique du niveau de complexité de dictées.

Quant à la seconde dimension, qui consiste à déterminer la complexité des différents phénomènes orthographiques et grammaticaux présents dans une dictée, on peut se reposer sur de nombreux travaux en linguistique et en pédagogie. Dans le cadre de cet article, l'accent est principalement mis sur les difficultés orthographiques décrites d'après le plurisystème graphique du français de Catach (1978). Son système organise, en cercles concentriques, les graphèmes de notre système graphique. Au centre se trouvent les archigraphèmes, qui sont transparents et simples à appréhender, mais plus on s'éloigne du centre, plus on rencontre des phénomènes complexes, qui nous intéressent plus particulièrement. On trouve ainsi les classes des morphogrammes (les marques de flexion et de dérivation), des logogrammes (représentations graphiques diverses pour une même prononciation et non justifiées par le système), et les lettres étymologiques et historiques. Le système de Nina Catach est toujours utilisé dans de nombreuses recherches aujourd'hui. Il est considéré comme l'un des modèles ayant permis de changer le regard sur l'orthographe, selon Angoujard (2001).

En parallèle, Nina Catach propose également une grille d'évaluation des difficultés qui se base sur son système (Catach, 1980). Elle organise les erreurs en plusieurs classes, allant des erreurs à dominante phonétique (cas où l'enfant ne maitrise pas encore la prononciation d'un mot, qu'il ne peut donc pas retranscrire correctement) jusqu'aux erreurs à dominante non fonctionnelle (associées aux lettres étymologiques et historiques). Ce sont les erreurs les plus complexes qui nous intéressent dans cet article.

3 Méthodologie

Dans cette section, nous présentons la façon dont le corpus de dictées servant pour l'entrainement du modèle statistique a été collecté (section 3.1). La section 3.2 liste ensuite les variables visant à capturer les difficultés de compréhension et les difficultés orthographiques et grammaticales des textes. Enfin, nous décrivons, à la section 3.3, le modèle statistique utilisé pour prédire le niveau de difficulté des dictées.

3.1 Collecte du corpus

Pour entrainer notre modèle prédisant la difficulté des dictées, il était nécessaire de disposer d'un corpus de dictées déjà classées par niveaux. Nous avons collecté des dictées depuis différents sites internet de référence qui peuvent être regroupées en deux catégories : les dictées de concours et les dictées d'apprentissage. Les dictées de concours ont été reprises de divers concours organisés dans la francophonie [1]. Bien que ces dictées risquent de modifier la cohérence des niveaux, en élevant sa complexité, il est intéressant de les conserver dans le corpus, car il s'agit d'une autre

1. À savoir les dictées du Balfroid, de la Fondation Paul Gérin-Lajoie, du Lions Club Gembloux, du championnat de Belgique, d'Orthosport, des Dicos d'or, du championnat suisse d'orthographe, du championnat du Cameroun d'orthographe, des Timbrés de l'orthographe, du Campus Eiffel, de la grande dictée Eric-Fournier et de la dictée des Amériques.

vision de l'orthographe et certaines difficultés peuvent être absentes des textes les plus simples et surreprésentées dans les plus complexes. Quant aux dictées d'apprentissage, ces dernières proviennent du Bescherelle, de manuels d'orthographe sur support informatisé, et peuvent parfois être ciblées sur une difficulté.

Niveau	Concours	Apprentissage
Primaire	88	114
Fin 3e secondaire	14	153
Fin 6e secondaire	28	125
Expert	83	35

TABLE 1 – Descriptif du corpus : nombre de textes par niveau et par type de dictées.

Ces dictées ont été classées en quatre catégories, en suivant les indications de niveau présentes sur chacun des sites consultés. Ces 4 catégories correspondent à des « niveaux » scolaires belges : (1) fin des primaires, (2) fin de la 3e secondaire, (3) fin de la 6e secondaire et (4) un niveau expert, qui correspond à des dictées plutôt destinées à un public universitaire. La table 1 décrit la distribution des dictées sur les 4 niveaux et en fonction de leur type (concours ou apprentissage).

3.2 Variables

Une fois le corpus constitué, nous nous sommes penchés sur l'identification des facteurs textuels susceptibles d'influencer (1) la bonne compréhension d'un texte et (2) sa réalisation orthographique.

Pour la première dimension, nous nous sommes basés sur les travaux en lisibilité et nous avons repris un ensemble de variables destinées à évaluer la complexité d'un texte à François (2011). Celles-ci sont surtout de deux types : des variables lexicales, comme la fréquence lexicale, la longueur des mots ou la densité du voisinage orthographique ; et des variables syntaxiques, incluant des informations sur la conjugaison, la longueur des unités syntaxiques et les ratios de catégories du discours. S'y ajoutent quelques variables sémantiques (cohésion moyenne interphrastique) et discursives (présence de dialogue), pour un total de 344 variables.

En ce qui concerne les variables spécifiques aux difficultés orthographiques et grammaticales typiques des dictées, nous nous sommes basés sur la littérature sur le sujet (*cf.* Section 2), qui a permis de mettre en avant cinq familles de variables à implémenter :

Les homophones : il s'agit de prendre en compte la densité des homophones, grammaticaux ou lexicaux, dans un texte. Plus celui-ci comporte d'homophones, plus il est susceptible d'être complexe, car la résolution de ces ambigüités mobilise davantage de ressources cognitives. Pour détecter les homophones, nous avons mis au point une liste d'homophones grammaticaux et une liste d'homophones lexicaux en nous basant sur diverses listes trouvées sur internet. Les critères retenus sont le nombre total d'homophones, le nombre d'homophones lexicaux, le nombre d'homophones grammaticaux, tous normalisés en fonction de la longueur du texte.

Les participes passés : les règles d'accord des participes passés sont connues pour être complexes en français. Les participes passés sont distingués selon leur « catégorie » (employé seul, avec être, avec avoir ou pronominal). De nouveau, plus un texte comprend de participes passés, plus sa difficulté augmente, en particulier si ces participes passés appartiennent aux catégories « employés avec avoir »

ou « avec un verbe pronominal ». Le tagger utilisé, TreeTagger (Schmid, 1994), ne précisant pas le type de participe passé, des règles ont été développées pour les distinguer. Sont considérés comme employés avec « avoir » ou « être », les participes passés précédés de ces deux auxiliaires dans les trois mots précédant le participe passé. Pour les participes passés pronominaux, les règles vérifient qu'avant le verbe « être », il y a bien deux pronoms (le sujet et la particule pronominale correspondante, ou la simple présence de « se »). Tous les autres cas sont considérés comme étant seuls.

Les terminaisons verbales homophoniques : les terminaisons verbales les plus fréquentes sont aussi les plus complexes, car plusieurs réalisations graphiques sont possibles pour un même phonème. Les terminaisons considérées sont celles en [e], [ɛ], [i], [õ] et [y]. La terminaison en [e] a d'ailleurs été considérée par Brissaud *et al.* (2006) comme une difficulté très importante pour l'apprentissage du français. Pour chacune des 5 terminaisons, nous avons calculé le ratio du nombre de verbes se terminant avec la terminaison sur le nombre total de verbes dans la dictée.

La correspondance phonème-graphème : dans un système d'écriture « simple », un phonème correspond à un seul graphème. C'est par ce stade que passe chaque enfant qui apprend l'alphabet. Malheureusement, le système du français est loin d'être aussi simple : à un phonème correspondent plusieurs graphies différentes. En nous reposant sur la typologie de Catach (1978), nous avons défini un ensemble de variables qui prennent en compte les graphèmes flexionnels du pluriel et du féminin, les doubles consonnes et les lettres étymologiques (y, h et ses dérivés). Le choix de ne modéliser que les graphèmes flexionnels et non tous les types de morphogrammes (tels que les dérivés, par exemple) est dû au fait qu'il s'agit sans doute des éléments qui présenteront les plus grosses difficultés, et qui sont aussi plus faciles à détecter.

La simplicité des mots : cette dernière variable s'intéresse aux mots considérés comme les plus simples de la langue française. Il s'agit d'une variable assez proche des variables de lisibilité reposant sur une liste de mots simples, décrites plus haut, mais qui adopte cette fois le point de vue de la production. Nous nous basons sur la liste de Nina Catach (Catach, 1984) [2] qui décrit, pour 3000 mots, l'ordre dans lequel les enfants sont supposés assimiler ces mots en production.

Au terme de cette étape, nous obtenons 344 variables « lisibilité » classées en 15 familles [3], ainsi que 31 variables « orthographiques », regroupées en 5 familles.

3.3 Notre modèle

Sur la base de ces variables, il a été possible de développer un modèle d'apprentissage automatique, en l'occurrence un séparateur à vastes marges (SVM) multiclasse et basé sur la stratégie « un contre un ». Nous avons tout d'abord étudié l'efficacité des variables en évaluant, pour chacune d'elles, sa corrélation de Pearson avec le niveau des dictées du corpus d'entrainement. Deux modèles ont dès lors été envisagés. D'une part, un modèle FULL, qui inclut toutes les variables « orthographiques », ainsi que la meilleure représentante de chaque famille des variables « lisibilité ». D'autre part, le modèle PARCIMONIEUX n'inclut, quant à lui, que les variables dont la corrélation de Pearson est significative en fonction d'un $\alpha = 0,01$. Cela revient à conserver 38 variables, 17 « orthographiques » et 21« lisibilité ». Nous avons également défini deux baselines : RANDOM, qui classe au hasard, et ONEFEAT, qui repose uniquement sur la meilleure variable : « ML3 », un modèle unigramme qui

2. La liste est celle proposée par Robert Rivière sur son site (http ://blog.orthodoc7.net/tag/Robert%20Rivi%C3%A8re), qui actualise et organise les mots en 5 classes.

3. Pour un détail des familles, se reporter à François (2011).

évalue la vraisemblance du texte à partir de la fréquence des formes fléchies désambigüisées.

Signalons que des expériences préliminaires ont montré qu'il était nécessaire de rééchantillonner les dictées afin que leur effectif par niveau soit égal, sans quoi le modèle obtenu favorisait trop la classe majoritaire. Un rééchantillonnage aléatoire stratifié a permis de retenir 118 textes par niveau, dont 48 ont été utilisés pour le corpus de développement et les 70 autres pour le corpus d'entrainement. Des tests ont été réalisés sur le corpus de développement en vue de déterminer les meilleurs métaparamètres des modèles SVM. Nous avons ainsi testés différents kernels (RBF, linéaire et polynomial), et avons effectué une *grid search* pour déterminer les valeurs optimales pour C et gamma, explorant l'espace entre 1000 et 0,001 [4]. Au terme de cette étape, un jeu optimal de métaparamètres a été retenu pour chacun des trois modèles (voir Table 2) dont les performances ont alors été estimées sur le corpus d'entrainement, via une procédure de validation croisée à dix échantillons.

4 Résultats

Les performances des trois modèles retenus au terme de l'étape de développement sont décrites à la table 2. Nous avons utilisé l'exactitude comme mesure d'évaluation principale – et comme fonction objective pour la sélection des modèles, mais l'erreur-type (RMSE) est également rapportée, car elle pénalise davantage les erreurs de prédiction les plus graves (ex. texte expert détecté comme primaire). RANDOM, qui classe au hasard, obtient une exactitude de 25%, tandis que ONEFEAT, basé sur la meilleure variable (« ML3 ») permet de classer correctement 37,85% des dictées. Le modèle PARCIMONIEUX, quant à lui, atteint une exactitude de 54% en validation croisée, contre près de 56% pour le modèle FULL. La sélection des variables semble donc ne pas être tout à fait optimale, puisqu'on perd près de 2% d'exactitude pour une réduction de seulement 16 variables. On notera cependant que la RMSE est quant à elle équivalente entre les deux modèles, ce qui signifie que le modèle parcimonieux ne commet pas plus d'erreurs graves.

Modèle	Nb. de variables	Kernel	C	γ	Exactitude	RMSE
RANDOM	0	/	/	/	25%	/
ONEFEAT	1	polynomial	100	/	41%	0,41
PARCIMONIEUX	38	RBF	1	0,5	54%	0,38
FULL	54	Linéaire	/	/	55,7%	0,38

TABLE 2 – Performances et caractéristiques des 4 modèles retenus.

Dans un second temps, nous avons analysé l'apport de chaque famille en entrainant, pour chacune des familles, le modèle FULL sans cette famille, ainsi qu'un modèle SVM ne contenant que les variables de cette famille. Les résultats obtenus sont résumés à la table 3. Au niveau des sous-familles, il apparait que ce sont les variables lexicales qui sont les plus efficaces, tant pour le jeu « lisibilité » (45,7%) qu'au niveau des variables orthographiques, où c'est la liste de Catach (1984) qui se distingue (47,8%). Les aspects syntaxiques ne viennent que bien après (40%) et sont suivis par la correspondance phonème-graphème (31,4%). Si on analyse l'effet de chacune des deux grandes familles, les variables

4. Les expérimentations ont été réalisées avec le logiciel Weka (Hall *et al.*, 2009).

orthographiques spécifiques à la dictée sont clairement moins prédictives (46,4%) que les variables évaluant la complexité de compréhension du texte (52,5%).

Famille de variables	Nb. var. incluses	Exac. sans famille	Exac. famille seule
FULL	54	/	55,7%
LISIBILITÉ.ALL	23	46,4	52,5 %
ORTHO.ALL	31	52,5 %	46,4 %
lexicale « lisibilité »	12	48,2%	45,7%
syntaxique « lisibilité »	7	54,6%	40%
sémantique/discursive « lisibilité »	4	55%	26%
homophones	3	55,7%	28,2 %
participes passés	9	55%	30,3%
terminaisons verbales homophoniques	6	56%	29,3%
correspondance phon.-graph.	7	54,3%	31,4%
simplicité lexicale	6	50%	47,8%

TABLE 3 – Analyse des variables : exactitude Performances et caractéristiques des 4 modèles retenus.

Les expériences d'ablation du modèle général (exactitude sans la famille) nous informent plus précisément sur la redondance informative de chacune des familles. Pour les variables orthographiques, on note que l'ablation de trois des cinq familles n'influence pas les performances du modèle, ce qui signifie que l'information qu'elles comportent est redondante avec d'autres variables. La correspondance phonème-graphème semble apporter une faible quantité d'information spécifique, tandis que la liste de Catach (1984) est quant à elle, de loin, la plus informative et cela, malgré la présence d'autres variables lexicales « lisibilité ». Cela semble confirmer qu'une liste de mots simples orientée vers la production apporte une information complémentaire par rapport à une liste orientée vers la réception [5].

5 Conclusion

Nous avons présenté un modèle prédisant automatiquement la difficulté de compréhension et orthographique de dictées qui repose sur un algorithme statistique par SVM. Il est capable de prédire correctement le niveau de dictées avec 56% d'exactitude, ce qui est en ligne, par exemple, avec les résultats obtenus par François & Fairon (2013) pour le français sur une tâche de lisibilité, relativement similaire. Ces performances restent toutefois trop faibles pour envisager tel quel une intégration du modèle dans une plateforme d'apprentissage de l'orthographe. Pour en améliorer les performances, il semble d'abord nécessaire de compléter le jeu de variables orthographiques et grammaticales afin de prendre en compte d'autres difficultés spécifiques aux dictées et ainsi augmenter l'information fournie au modèle. En effet, la majorité de nos variables spécifiques au contexte de la dictée se sont révélées décevantes, à l'exception de celles basées sur la liste de Catach (1984).

Par ailleurs, il serait intéressant d'étudier plus en détails l'homogénéité des annotations du corpus sur le modèle de l'analyse de François (2014). En effet, il peut y avoir un impact du type de dictées (apprentissage ou concours) sur l'entrainement du modèle et, de ce fait, il serait bon de déterminer si se limiter aux dictées d'apprentissage ne permettrait pas de disposer d'un corpus moins bruité.

5. Dans ce cas, la liste utilisée dans François (2011) est celle de Gougenheim *et al.* (1964).

Références

ANGOUJARD A. (2001). L'orthographe à l'école : héritages et perspectives. In *Lettres ouvertes aux enseignants de français. L'orthographe : accords et désaccords*, volume 14-15 : CRDP de Bretagne.

BEAUFORT R. & ROEKHAUT S. (2011). Le tal au service de l'alao/elao. l'exemple des exercices de dictée automatisés. In *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2011)*, p. 87–92.

BOURDIN B. (1994). *Cout cognitif de la production verbale. Etude comparative oral ecrit chez l'enfant et l'adulte*. PhD thesis, Dijon.

BRISSAUD C., CHEVROT J.-P. & LEFRANÇOIS P. (2006). Les formes verbales homophones en/e/entre 8 et 15 ans : contraintes et conflits dans la construction des savoirs sur une difficulté orthographique majeure du français. *Langue française*, (3), 74–93.

BRISSAUD C. & MORTAMET C. (2015). Présentation. *Glottopol*, (26), 2–10.

CATACH N. (1978). *L'orthographe*. Paris : PUF.

CATACH N. (1980). *L'orthographe française*. Paris : Nathan.

CATACH N. (1984). *Les listes orthographiques de base du français*. Paris : Nathan.

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.

DALE E. & CHALL J. (1948). A formula for predicting readability. *Educational research bulletin*, **27**(1), 11–28.

DASCALU M. (2014). Readerbench (2)-individual assessment through reading strategies and textual complexity. In *Analyzing Discourse and Text Complexity for Learning and Collaborating*, p. 161–188. Springer.

DELABARRE É. & DEVILLERS M.-L. (2015). Mise(s) en œuvre d'une activité orthographique : la dictée. *Glottopol*, (26), 48–57.

FENG L., JANSCHE M., HUENERFAUTH M. & ELHADAD N. (2010). A Comparison of Features for Automatic Readability Assessment. In *COLING 2010 : Poster Volume*, p. 276–284.

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

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

FRANÇOIS T. & FAIRON C. (2013). Les apports du TAL à la lisibilité du français langue étrangère. *Traitement Automatique des Langues (TAL),*, **54**(1), 171–202.

FRANÇOIS T. (2014). An analysis of a french as a foreign language corpus for readability assessment. In *Proceedings of the 3rd workshop on NLP for Computer-assisted Language Learning, NEALT Proceedings Series Vol. 22, Linköping Electronic Conference Proceedings 107*, p. 13–32.

GOUGENHEIM G., MICHÉA R., RIVENC P. & SAUVAGEOT A. (1964). *L'élaboration du français fondamental (1er degré)*. Paris : Didier.

HALL M., FRANK E., HOLMES G., PFAHRINGER B., REUTEMANN P. & WITTEN I. H. (2009). The weka data mining software : an update. *ACM SIGKDD explorations newsletter*, **11**(1), 10–18.

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

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

RUGGIA S. (2000). La dictée interactive. *Apprentissage des langues et systèmes d'information et de communication*, **3**(1).

SANTIAGO ORIOLA C. (1998). *Système vocal interactif pour l'apprentissage des langues. La synthèse de la parole au service de la dictée*. PhD thesis, Université de Toulouse 3.

SCHELSTRAETE M.-A. & MAILLART C. (2004). Développement des mécanismes orthographiques et limitations de traitement. *Glossa*, (89), 4–20.

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.

VAJJALA S. & MEURERS D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, p. 163–173.