Ma copie adore le vélo : analyse des besoins réels en correction orthographique sur un corpus de dictées d'enfants

Jean-Yves Antoine, Marion Crochetet, Celine Arbizu, Emmanuelle Lopez, Samuel Pouplin, Amélie Besnier, Mathieu Thebaud

▶ **To cite this version:**

Jean-Yves Antoine, Marion Crochetet, Celine Arbizu, Emmanuelle Lopez, Samuel Pouplin, et al.. Ma copie adore le vélo : analyse des besoins réels en correction orthographique sur un corpus de dictées d'enfants. TALN 2019, Jun 2019, Toulouse, France. hal-02375246

HAL Id: hal-02375246
https://hal.science/hal-02375246

Submitted on 21 Nov 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Ma copie adore le vélo : analyse des besoins réels en correction orthographique sur un corpus de dictées d'enfants

Jean-Yves Antoine[1], Marion Crochetet, Céline Arbizu, Emmanuelle Lopez[2], Samuel Pouplin[3], Amélie Besnier[4], Mathieu Thebaud[4]
(1) LIFAT, ICVL, 41000 Blois, France
(2) CRTLA, Hôpital R. Poincaré, 92380 Garches France
(3) PFNT, Hôpital R. Poincaré, 92380 Garches, France
(4) CMRRF de Kerpape, Ploemeur, 56270 France

Jean-yves.antoine@univ-tours.fr, marion.crochetet@aphp.fr, celine.arbizu@aphp.fr, emmanuelle.lopez@aphp.fr, abesnier@kerpape.mutualite56.fr, mthebaud@kerpape.mutualite56.fr

RÉSUMÉ

Cet article présente la constitution d'un corpus de textes produits, sur des données lors de dictées, par des enfants paralysés cérébraux (PC) ou dysorthographiques, son annotation en termes d'erreurs orthographiques, et enfin son analyse quantitative. Cette analyse de corpus a pour objectif de définir des besoins réels en matière de correction orthographique, et ce pour les personnes souffrant de troubles du langage écrit comme pour le grand public. Notre étude suggère que les correcteurs orthographiques ne répondent que partiellement à ces besoins.

ABSTRACT

A corpus analysis to define the needs of dyslexic children in terms of spelling correction

This paper presents a corpus of texts produced, during dictations, by dyslexic children, its annotation in terms of spelling errors, and finally its quantitative analysis. The purpose of this corpus analysis is to define real needs concerning automatic spelling correction, both for people suffering from language disorders and for ordinary people. Our study suggests that current spell checkers are unable to meet the majority of these needs.

MOTS-CLÉS : Correction orthographique, dyslexie, apprentissage des langues, analyse corpus

KEYWORDS: Spelling correction, dyslexia, language learning, corpus analysis

1 Introduction

La correction orthographique est une des applications les plus anciennes du TAL, puisque les premières recherches du domaine datent des années 1960s (Damereau 1964). Elle constitue une application emblématique du fait du rôle central de la maitrise d'une langue dans la représentation sociale d'un individu. Pourtant, si l'on en juge par sa faible représentation dans la littérature récente, la correction orthographique n'est plus considérée comme une problématique d'actualité. Est-ce à

dire que le verrou scientifique posé par la correction automatique est résolu? Nous pouvons en douter lorsque nous observons que les systèmes de correction automatique du commerce, mais également ceux issus de la recherche académique, peinent à détecter des erreurs orthographiques qui ne conduisent pas à la production d'un non-mot (cf § 4).

Dans cet article, nous abordons la question de la correction automatique dans un cadre particulier : celui des productions d'enfants apprenants en situation de handicap (paralysie cérébrale d'une part, dysorthographie d'autre part). La correction orthographique est destinée ici à être couplée à la prédiction de mots d'une système d'aide à la communication (Wandmacher et al. 2007). L'objectif n'est donc pas de corriger des énoncés complets, mais de prédire à la volée la suite d'une séquence de lettres qui est potentiellement erronée. Ainsi, si l'utilisateur a saisi le début de phrase *un ba..*, le système doit pouvoir prédire le mot *bateau*, mais également l'adjectif *beau,* pour corriger une éventuelle erreur d'encodage graphémique du phonème \o\. La combinaison correction/prédiction, explorée par (Li et al. 2013) est rarement envisagée, en particulier dans le cas de la dyslexie.

La tâche est donc plus complexe de celle envisagée pour les correcteurs orthographiques classiques. Pour mieux l'appréhender, nous avons mené une analyse des besoins sur des données issues de dictées réalisées par des enfants suivis au centre de rééducation de Kerpape (Mutualité du Morbihan) et à l'Hôpital Raymond Poincaré de Garches. L'annotation des erreurs orthographiques contenues dans le corpus, de même que l'analyse du comportement de correcteurs orthographiques sur cette ressource, nous a permis de dresser un ensemble de besoins qui sont autant de défis encore mal résolus. Cette étude reste préliminaire en termes de représentativité du corpus traité. Ses conclusions nous semblent toutefois assez éclairantes pour rouvrir la question d'une correction approfondie intégrée à la prédiction de mots.

Après un bref état de l'art sur la question, nous présentons le recueil et l'annotation du corpus sur lequel nous avons travaillé. Enfin, nous détaillons les résultats de l'analyse de besoins menée sur le corpus. En conclusion, nous esquissons une stratégie de correction qui sera mise en œuvre dans le cadre d'un projet, PREDICT4ALL, financé par la fondation Bennetot, fondation de la MATMUT.

2 Correction et erreurs orthographiques

2.1 Typologie des erreurs orthographiques

Les erreurs orthographiques qui surviennent dans un texte sont de nature variées. Plusieurs typologies d'erreurs ont été proposées dans la littérature, qui diffèrent par leurs objectifs. Certaines cherchent à rendre compte du comportement cognitif du scripteur et sont utiles à l'orthophoniste pour faire un bilan qui permettra l'adaptation du correcteur à son utilisateur. D'autres se concentrent uniquement sur la forme qu'aura à traiter le correcteur. (Kukich 1992) distingue ainsi :

- Les erreurs lexicales, qui conduisent à un non mot (mot absent du dictionnaire).
- Les erreurs syntaxiques qui conduisent à une phrase agrammaticale : par exemple, la partie du discours du mot n'est pas celle attendue, ou l'accord entre les mots n'est pas assuré.
- Les erreurs sémantiques, qui conduisent à une phrase incohérente sémantiquement.

Dans le cas des erreurs syntaxiques et sémantiques, le mot erroné mal orthographié correspond à une entrée du dictionnaire (*real-word errors*), ce qui rend leur détection plus délicate.

2.2 Correction orthographique : un bref état de l'art

Correction hors contexte : distance d'édition – Originellement, la correction automatique s'est focalisée sur les erreurs lexicales. Leur détection est immédiate, il reste à corriger le mot en cherchant dans le dictionnaire le mot le plus proche suivant un modèle d'erreur donné. Des modèles d'erreurs classiques sont la distance d'édition de Levensthein (1966), ou des modèles phonétiques comme Aspell (Atkinson 2011). La limite bien établie de ces approches est que les performances se dégradent lorsqu'on augmente la taille du dictionnaire: il y a de plus en plus de formes acceptables, ce qui fait qu'une erreur orthographique peut correspondre facilement à une autre forme lexicale : on se trouve alors dans le problème central de la détection des erreurs syntaxiques ou sémantiques.

Correction contextuelle : ensemble de confusion – Pour traiter les erreurs syntaxiques ou sémantiques, il est nécessaire de considérer le contexte d'occurrence du mot. Celui-ci n'est alors jugé correct que si sa probabilité d'apparition est supérieure à celle des mots qui lui sont proches orthographiquement, au sens du modèle d'erreur utilisé par l'approche hors-contexte. Chaque mot est ainsi associé à un ensemble de confusion (par exemple : *désert*, *dessert*) au sein duquel on cherche les termes de probabilité maximale suivant le modèle du canal bruité (Golding & Roth 1999, Norvig 2009). Le risque principal est de proposer une correction lorsque le mot est correct. Pour cette raison, les correcteurs privilégient la précision au rappel. Pour limiter la sur-correction, certains proposent de définir des listes noires de termes à ne pas corriger (Whitelaw et al. 2009), d'autres favorisent le mot saisi en n'envisageant son remplacement que si l'accroissement de probabilité dépasse un certain seuil (Jurafsky, Martin 2018). Le risque de sur-correction conduit également les correcteurs à définir les ensembles de confusion à une distance d'édition de 1 (erreurs simples). Ce choix se base sur les observations selon lesquelles 80% des erreurs orthographiques correspondent à une distance d'édition de 1 (Damereau 1964 ; Pollock et Zamora 1984). Un des enjeux de notre étude de corpus sera de savoir si cette conclusion faite sur l'anglais reste pertinente en français.

Correction sémantique –Enfin, certains auteurs envisagent une correction purement sémantique, en modélisant le contexte thématique d'apparition d'un mot par plongements de mots (Nagataa, 2017). Cette approche montre sa limite sur les langues à la morphologie suffisamment riche, pour lesquelles l'erreur peut être de flexion et ne pas conduire à un changement de lemme, donc de sémantique.

3 Corpus PREDICT4ALL d'erreurs orthographiques

L'objectif du projet PREDICT4ALL est de développer un moteur de prédiction de mots, utilisé dans un contexte d'aide à la communication orale ou écrite, qui s'adapte à des personnes en situation de handicap moteur et à des personnes présentant des troubles dysorthographiques ou simplement d'apprenants du français. Ces systèmes, tels le système Sibylle (Wandmacher et al. 2007), reposent sur une prédiction lexicale qui est perturbée par les erreurs orthographiques. Notre ambition est de combiner prédiction et correction à la volée. Les travaux sur la correction orthographique se sont rarement penchés sur le problème de l'apprentissage ou de la dyslexie. L'état de l'art récent se limite à notre connaissance à (Pedler 2007) et (Rello et al. 2015), qui ne portent pas sur le français.

Le corpus PREDICT4ALL réunit des données obtenues lors de dictées réalisées par 10 enfants scolarisés en CM (Cours Moyen de l'école primaire). A des fins de comparaison, les dictées étaient identiques et extraites de (Baneath 2015). Deux populations ont été distinguées :

- **DYS : Dyslexiques** – 5 enfants dysorthographiques (âge moyen 10 ans), diagnostiqués et suivis au centre de référence des troubles du langage de l'hôpital de Garches. Nous ne détaillons pas leur tableau clinique du fait de la taille réduite de la cohorte considérée.

– **PC** – 5 enfants paralysés cérébraux sans troubles langagiers, suivis au Centre Mutualiste de Rééducation et de Réadaptation Fonctionnelle de Kerpape. Ces apprenants souffrent de retards cognitifs qui ont un impact sur leur âge langagier. Plus qu'un groupe contrôle, ce public est la seconde cible visée par le projet PREDICT4ALL.

Ce corpus réunit 521 erreurs orthographiques. En dépit de sa taille limitée, seules deux autres ressources peuvent être comparées à la nôtre : d'une part un corpus anglais de 2654 erreurs mélangeant des sources assez hétérogènes (Pedler 2007) et d'autre part un corpus espagnol de 1171 erreurs (Rello and al. 2014). (Plisson et al. 2013) proposent un corpus de français canadien avec des enfants dysorthographiques en production libre et non pas en activité contrôlée de dictée. Ce corpus a l'avantage de la naturalité, mais ne peut permettre des comparaisons non biaisées entre scripteurs. En l'état actuel, la taille réduite de notre corpus interdit une caractérisation statistique du lien qui peut exister entre profil dyslexique et comportement langagier. Notre étude sera ainsi complétée, dans l'année à venir, sur une ressource de plus grande envergure en cours de constitution. L'étude pilote que nous présentons ici nous semble toutefois déjà présenter des résultats dignes d'intérêt.

Nous avons adopté un schéma d'annotation qui répond à deux besoins distincts : d'une part celui des chercheurs en TAL qui utilisent différentes ressources (lexiques, parseurs) pour analyser chaque chaîne de caractères. D'autre part, celui des orthophonistes qui étudient le comportement langagier de l'utilisateur, pour adapter au mieux la correction à son tableau clinique. Chaque erreur est caractérisée par les traits d'annotation suivants :

– **Erreur distincte** (TAL) – Précise si un mot comporte une ou plusieurs erreurs différentes. Plusieurs annotations distinctes pourront donc être distinguées dans un mot.
– **Type d'erreur** (TAL) – Lexicale / syntaxique / sémantique, inspirée de (Kukich 1992).
– **Morphologie** (TAL) **–** Impact de l'erreur en termes de forme, inspirée de (Damerau 1964). Nous distinguons d'une part les erreurs de segmentation : fusion (exemple : *lecole*) ou séparation de mots (exemple : *mon tagne*). D'autre part, en l'absence d'erreur de segmentation, nous comptabilisons la distance d'édition entre le mot écrit et le mot attendu.
– **Phonologie** (orthophonie) **–** Nous avons bâti une typologie basée sur la phonologie inspirée de (Catach 1980) tout en utilisant la terminologie erreurs phonologiquement plausibles (EPP) / non plausibles (ENPP) de (Martinet et Valdois, 1999). Celle-ci est utilisée en usage clinique par les orthophonistes pour repérer des compétences phonologiques déficitaires. Parmi les EPP, nous caractérisons les cas d'homophonie, le mauvais choix de graphème pour rendre un phonème (*le chamo*), les erreurs sur lettres muettes (*la souri*), les liaisons erronées (*dans sune des poches*), le mauvais encodage des semi-voyelles (*voillait*), les erreurs de flexion (*tu mange*) et celles portant sur des orthographes irrégulières (*fame* vs. *femme*). Pour les ENPP, nous distinguons les cas où l'erreur porte sur une graphie contextuelle, c'est-à-dire une graphie dont la prononciation dépend des voyelles environnantes (exemple : *gide* pour *guide*), les substitutions phonétiques, dont celles correspondant au changement d'un trait acoustique (voisement par exemple), les omissions ou insertions de graphèmes non muets, les erreurs séquentielles (*délcare* vs. *déclare*) et enfin les paragraphies morphémiques caractérisées, soit la production d'un mot qui partage avec le mot cible sa racine mais s'en différencie par un affixe (exemple : *ces*/*cette*, *un*/*une*...)

Enfin, nous avons étudié le comportement de plusieurs correcteurs grand public sur le corpus : le correcteur de *Microsoft Word 10* (2016), le correcteur en ligne *LanguageTool* (`languagetool.org/fr`) et enfin le correcteur *Cordial* de Synapse (version 11, 2005) qui présente l'intérêt de mettre en jeu une correction contextuelle avec analyse syntaxique profonde. Pour chaque erreur, nous distinguons les cas où (1) l'erreur se trouve parmi les propositions de correction, (2) l'erreur est détectée mais les propositions de correction sont erronées et enfin (3) l'erreur n'est pas détectée. La section suivante présente les analyses quantitatives réalisée sur cette ressource.

4 Analyse des besoins : étude quantitative du corpus

Répartition des erreurs – La distribution des erreurs permet de caractériser les besoins qui existent en matière de correction orthographique, pour les enfants PC et les enfants dysorthographiques.

Sous-corpus	Nb. de mots	Nb. mots erronés	**Taux de mots erronés**	Nb. total d'erreurs	**Nb. d'erreur par mot erroné**	**% mots avec erreur multiple**
PC	415	120	**29 %**	152	**1,3**	**20 %**
DYS	415	227	**55 %**	409	**1,8**	**45 %**

TABLE 1 : Distribution globale des erreurs dans le corpus PREDICT4ALL

La table 1 présente ces distributions sur les sous-corpus PC et DYS. Comme attendu, les troubles DYS rendent plus difficile la tâche de la correction. Nous observons comme (Plisson et al. 2013) une augmentation de la fréquence des erreurs chez les personnes DYS. Au final, plus de la moitié des mots (55%) du corpus DYS sont erronés. Ces erreurs sont par ailleurs plus profondes. Le taux moyen d'erreurs dans un mot erroné passe ainsi de 1,3 (PC) à 1,8 (DYS) : près de la moitié des mots erronés contiennent plusieurs erreurs chez les personnes DYS. Dans l'énoncé *il se défande contre les mousique*, les mots ***défande* et ***mousique* combinent une erreur de flexion avec respectivement un mauvais graphème ou une omission. Cette accumulation a un impact sur le type d'erreur.

Type d'erreur – Comme le montre la table 2, la proportion d'erreurs non lexicales se réduit ainsi de 49% à 29% entre les sous-corpus PC et DYS : l'accumulation des erreurs commises par les patients DYS réduit les chances de produire un mot du dictionnaire. Cette situation pourrait être un avantage pour la correction automatique, puisque ces erreurs sont détectables sans prise en compte du contexte. On observe toutefois que l'augmentation de la proportion d'erreurs lexicales est due pour partie à celle des erreurs de segmentation[1]. Or, la plupart de ces erreurs constituent un véritable défi pour la correction (par exemple : *janga jerer* pour de *j'engagerai*). Notons que ces problèmes de segmentation se retrouvent aussi chez les apprenants PC (9 % des cas). (Plisson et al. 2013) observe une même évolution (passage de 8% à 12% des cas) sur le français canadien.

Sous-corpus	Erreurs syntaxiques	Erreurs sémantiques	**Total erreurs non lexicales**	Erreurs lexicales (hors segmentation)	*Erreurs de segmentation*
PC	49%	0%	**49%**	42%	*9%*
DYS	26%	3%	**29%**	56%	*15%*

TABLE 2 : Distribution des erreurs du corpus PREDICT4ALL par type

Dans les deux populations, on observe enfin que le taux d'erreurs non lexicales est significativement supérieur à celui observé sur d'autres langues : 9% en espagnol chez (Rello et al. 2014) et 17% en anglais chez Pedler (2007). Une explication raisonnable à ce particularisme réside dans la forte morphologie flexionnelle du français. Ainsi, 47% des erreurs du sous-corpus PC concernent la flexion (table 4) et conduisent pour la plupart à des erreurs non lexicales (exemples : *explorer*, *explorai* et *exploré*) dont la détection nécessite une analyse syntaxique, même locale.

Distance d'édition – L'hypothèse selon laquelle ¾ des erreurs orthographiques sont à une distance d'édition de 1 (Damerau 1964) est retrouvée par (Rello et al. 2014) avec des personnes dyslexiques hispanophones (73% d'erreurs simples). Nos résultats ne confirment pas ces observations sur le

[1] Rares sont les erreurs de segmentation donnant des mots lexicaux, tel *ma copine a dore le vélo*

français (Table 3): les erreurs simples ne représentent que la moitié environ des situations dans nos deux populations. Dans près d'un quart des cas, la distance d'édition est même strictement supérieure à 2. Nos résultats recoupent ceux de (Pedler 2007) et (Baeza-Yates et Rello 2011) qui ne trouvent respectivement que 58% et 53% d'erreur simples sur l'anglais. Il s'agit d'un argument fort contre la définition d'ensembles de confusion limités à une distance d'édition de 1.

Sous-corpus	Distance 1	Distance 2	Distance > 2	**Total erreurs multiples**	*Erreur sur 1° caractères*
PC	52 %	26 %	22 %	**48 %**	*4 %*
DYS	46 %	29 %	25 %	**54 %**	*14 %*

TABLE 3 : Distribution des erreurs en fonction de la position et de la distance d'édition

Enfin, notons que le taux d'erreurs portant sur la première lettre du mot varie entre 4 % (corpus PC) et 14 % (DYS). Ces ordres de valeur assez modestes se retrouvent sur l'anglais (Pedler 2007, Pollock et Zamora 1984, Yannakoudakis et Fawthrop 1983). Dans le cadre du projet PREDICT4ALL, la correction est couplée avec une prédiction lexicale. Le faible taux d'erreurs en début de mots nous dissuade de tenter une correction lorsque seule la première lettre du mot est déjà saisie. Trop de faux positifs seraient en effet prédits. Les corrections sur la première lettre seront envisagées uniquement lorsque 2 lettres auront été entrées par l'utilisateur.

Phonologie – La sévérité des erreurs observées chez les personnes DYS (erreurs de segmentation, distance d'édition élevée) incite à proposer une correction adaptée à l'utilisateur et non à envisager une correction générique. Cette adaptation s'appuiera sur le bilan, réalisé en orthophonie, des types d'erreurs systématiques de l'utilisateur. Nous ne détaillerons pas la répartition des types d'erreurs observées dans le corpus. La table 4 nous suggère toutefois que les troubles DYS augmentent le risque de produire des erreurs qui ne sont pas phonologiquement plausibles, observation déjà relevée par (Plisson et al. 2013), qui observe par ailleurs un même niveau général d'EPP.

Sous corpus	**Phonologiquement plausible**				**Phonologiquement non plausibles**				
	Total	*Graphème erroné*	*Flexion*	*Autre*	**Total**	*Substitut. phonétique*	*Insertion Omission*	*Séquence*	*Autre*
PC	**71 %**	*12 %*	*47 %*	*12 %*	**29 %**	*4 %*	*14 %*	*1 %*	*10 %*
DYS	**62 %**	*18 %*	*23 %*	*21 %*	**38 %**	*13 %*	*15 %*	*4 %*	*6 %*

TABLE 4 : Distribution des erreurs en fonction de la position et de la distance d'édition

Les erreurs séquentielles (exemple : **graçon*) et les substitutions phonétiques (**crande* vs *grande*), qui sont marginales sur le corpus PC, représentent 17 % des erreurs des personnes DYS. Leur fréquence est toutefois variable d'une personne à l'autre et pourrait dépendre du type de dysorthographie. Le recensement de ces erreurs dans le corpus nous a toutefois permis d'obtenir des patrons de correction systématiques. Il nous semble ainsi important de définir, comme modèle d'erreur, des distances d'édition adaptées au tableau clinique de chaque personne.

5 Analyse des besoins : étude de performances

La correction que nous envisageons est réalisée à la volée en cours de saisie, ce que ne font pas les correcteurs standards. Nous avons toutefois examiné leur efficacité sur nos écrits d'enfants PC ou

dyslexiques. Les résultats (Table 5) mettent en évidence certaines limites des correcteurs, qui sont dans l'incapacité de corriger plus de la moitié des erreurs du sous-corpus PC. Dans plus d'un tiers des cas, ils ne détectent même pas ces erreurs, avant tout lorsqu'il s'agit d'erreurs non lexicales. Les taux de non-correction montent jusqu'aux trois quarts sur le sous-corpus DYS. Nos observations corroborent celles de (Pedler 2007) menées sur des correcteurs de l'anglais. On remarque enfin que mieux un correcteur se comporte sur le corpus PC, moins il s'en sort sur le corpus DYS. De fait, alors que *Cordial* corrige des erreurs syntaxiques négligées par les autres correcteurs sur le corpus PC, son parseur semble fortement désorienté sur les écrits DYS.

Corpus	**PC**			**DYS**		
Erreur	Corrigé	Détecté	Non détecté	Corrigé	Détecté	Non détecté
LanguageTool	**36 %**	28 %	36 %	**26 %**	43 %	31 %
Word	**42 %**	19 %	39 %	**24 %**	40 %	36 %
Cordial	**44 %**	19 %	37 %	**20 %**	39 %	41 %

TABLE 5 : Performances des correcteurs orthographiques sur le corpus

6 Conclusion

Notre étude, originale sur le français, pose ou retrouve plusieurs recommandations pour une correction adaptée à des enfants présentant ou non une dysorthographie :

- La correction doit être envisagée à des distances d'édition de 2 ou 3 (et non 1),
- La correction sera adaptée à l'utilisateur, après bilan en orthophonie : on évitera les distances d'édition génériques au profit de distances paramétrables relevant de types d'erreurs précis,
- La correction ne doit pas être envisagée dès la saisie de la première lettre d'un mot,
- Les erreurs de segmentation ne peuvent être ignorées,
- Une correction contextuelle est essentielle au vu de l'importance des erreurs non lexicales.

Partant de cette analyse, nous avons développé un prototype de correction qui est intégré à la prédiction du système SIBYLLE. La prédiction est lancée à la fois sur les séquences saisies par l'utilisateur, mais également sur les chaînes corrigées respectant les contraintes ci-dessus. L'évaluation d'un premier prototype sur un extrait du corpus *Le Monde* montre que le taux d'économie de saisie de la prédiction (KSR-5) ne décroit que légèrement (47% contre 54% sans correction). La correction ne génère donc pas trop de faux positifs sur les énoncés corrects. Ses capacités de correction ont été testées sur le corpus d'erreurs WiCoPaCo (Max et Wisniewski 2010) issu de pages de révision de Wikipedia. Alors que la prédiction permet de corriger 69% des mots du corpus (dès qu'une bonne prédiction est proposée en cours de saisie, on choisit cette solution, évitant de fait certaines erreurs à venir), l'ajout de la correction élève ce taux à 74%. Ces premiers résultats ont été obtenus sur un prototype encore non optimisé. Ils sont toutefois suffisamment encourageants pour suggérer la pertinence d'une combinaison entre correction et prédiction dans le cadre de l'aide à la communication dédiée à des enfants PC ou dyslexiques.

Remerciements

Cette recherche a été financée par la Fondation Bennetot, Fondation de la Matmut, dans le cadre du projet PREDICT4ALL.

Références

Atkinson K. (2011). Gnu aspell.

Baeza-Yates R., Rello L. (2011) Estimating dyslexia in the web. Proc. Int. Cross-Displinary Conference on Web Accessibility, W4A'2011. Vol. 8, 1-4.

Baneath B., Alberti C., Boutard C., Gatignol P. (2015). *Chronodictées, outils d'évaluation des performances orthographiques*. Ortho-Editions ;

Catach N. (1980) *L'orthographe française: traité théorique et pratique avec des travaux d'application et leurs corrigés*. Coll. Nathan Université, Paris, Nathan.

Damerau F. (1964). A technique for computer detection and correction of spelling errors. *Communications of the A.C.M.*, 7:171–176

Golding A.R., Roth D. (1999). A Winnow based approach to context-sensitive spelling correction. *Machine Learning*, 34(1-3), 107–130

Jurafsky D., Martin J.H. (2018). Spelling erros correction and the noisy channel. In Jurafsky D., Martin J.H. *Speech & Language Processing* (3rd. ed.), Prentice Hall, Pearsons Ed. Apprendix B.

Kukich K. (1992). Techniques for automatically correcting words in text. *ACM Computing Survey*s, 24(4), 377–439

Lefavrais P. (2006). *Test de leximétrie de l'Alouette*. Paris: ECPA

Levensthein V. (1966) Binary codes capable of correcting deletions, insertions and reversions. *Soviet Physics Dokkady* 10, 845-848 (trad.)

Li A.Q., Sbattella L., Tedesco R. (2013) PoliSpell : an adaptive spellchecker and predictor for people with dyslexia. Proc. UMAP 2013. In Carberry S. et al. (Eds.) LNCS 7899, 302-309.

Martinet C., Valdois S. (1999). L'apprentissage de l'orthographe d'usage et ses troubles dans la dyslexie développementale de surface. *L'Année psychologique*, 99(4), 577-622

Max A., and Wisniewski G. (2010). Mining Naturally-occurring Corrections and Paraphrases from Wikipedia's Revision History, Actes *LREC 2010*. La Valette, Malte.

Nagataa R., Takamurab H., Neubigc G. (2017). Adaptive spelling error correction models for learner English. Actes de *KES'2017*, Marseille.

Norvig P. (2009). Natural language corpus data. In Segaran, T. and Hammerbacher, J. (Eds.), *Beautiful data: the stories behind elegant data solutions*. O'Reilly.

Pedler J. (2007). *Computer Correction of Real-word Spelling Errors in Dyslexic Texts*. PhD Thesis. London University.

Plisson A., Daigle D., Montésinos-Gelet I. (2013). The spelling skills of French-speaking dyslexic children. *Dyslexia*, 1 :76-91.

Pollock J.J., Zamora A. (1984). Automatic Spelling Correction in Scientific and Scholarly Text. *Communications of the A.C.M*. 27 (4): 358-68.

Rello R., Baeza-Yates R., Llisterri J. (2014). DysList: An Annotated Resource of Dyslexic Errors. Actes *LREC'2014*. 1289-1296.

Rello R., Ballesteros M., Bigham J.P. (2015). A Spellchecker for Dyslexia. Actes *ACM TASSETS'2015*.

Wandmacher T., Antoine J.-Y., Poirier F. (2007) Sibylle : a system for alternative communication adapting to the context and its user. Actes *ACM Conference on Assistive Technologies. ASSETS'2007*, Phoenix. Arizona. 203-210.

Whitelaw C., Hutchinson B., Chung G.Y., Ellis G. (2009). Using the web for language independent spellchecking and autocorrection. Actes de *EMNLP'2009*, 890–899.

Yannakoudakis, E.J., Fawthrop,D. (1983) The Rules of Spelling Errors. *Information Processing and Management.* 19 (2), 87-99.