Spécificités des erreurs d'orthographe des personnes dyslexiques : analyse d'un corpus de productions écrites

Johana Bodard

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Spécificités des erreurs d'orthographe des personnes dyslexiques : analyse d'un corpus de productions écrites

Johana Bodard
Laboratoire CHArt, 2 rue de la Liberté, Saint-Denis, France
johana.bodard@etud.univ-paris8.fr

RÉSUMÉ

Cet article présente un travail d'analyse des erreurs d'orthographe de personnes dyslexiques à partir de corpus écrits en langue française. L'objectif de cette analyse est d'étudier la fréquence et les caractéristiques des erreurs afin de guider le développement de modules de correction orthographique spécifiques. Les résultats de cette analyse sont comparés aux travaux déjà réalisés en français, anglais et espagnol.

ABSTRACT

What are the characteristics of spelling errors made by dyslexics: an analysis of errors based on written corpora

In this paper, we present an analysis of spelling errors made by French dyslexics based on written corpora. The objective of this analysis is to investigate the frequency and characteristics of the spelling errors in order to guide the development of specific spell checking modules. The results of this analysis are compared with similar works in French, English and Spanish.

MOTS-CLÉS : dyslexie, analyse de corpus, correction orthographique.

KEYWORDS: dyslexia, corpus analysis, spell checking.

1 Introduction

La dyslexie est un trouble spécifique des apprentissages affectant le langage écrit dont la prévalence en France est estimée entre 6 et 8 % (Barrouillet *et al.*, 2007). Ce trouble entraine notamment des difficultés importantes dans l'acquisition de l'orthographe (dysorthographie), difficultés qui persistent souvent à l'âge adulte (Mazur-Palandre, 2018). Le correcteur orthographique apparait comme un outil particulièrement adapté pour pallier les difficultés orthographiques des personnes dyslexiques. Cependant, les correcteurs orthographiques classiques s'avèrent peu performants sur les écrits des dyslexiques (Bacquelé, 2015; Antoine *et al.*, 2019). Parmi les hypothèses avancées pour expliquer les faibles performances des correcteurs classiques sur ce type d'écrits, on peut citer : l'impossibilité de certains dyslexiques d'écrire correctement les initiales des mots (Bacquelé, 2015), le nombre important d'erreurs produisant des mots présents dans le dictionnaire (Antoine *et al.*, 2019), le nombre élevé d'erreurs par mot (Antoine *et al.*, 2019), la présence de mots mal découpés (fusionnés ou fragmentés) (Antoine *et al.*, 2019; Sitbon *et al.*, 2007), une écriture fortement phonétique (Sitbon *et al.*, 2007).

Afin de mettre en place des algorithmes de correction orthographique adaptés aux écrits des personnes

Volume 3 : Rencontre des Étudiants Chercheurs en Informatique pour le TAL, pages 15–28.
HAL : hal-02786183.

dyslexiques, nous avons réalisé un travail préalable d'analyse des erreurs d'orthographe à partir de corpus écrits. Nous avons extrait et annoté les erreurs d'orthographe afin de vérifier les hypothèses citées dans le paragraphe précédent et de comparer les résultats obtenus à ceux des quelques travaux existant sur le sujet. Notre objectif est de guider les choix algorithmiques que nous opérerons lors du développement des modules de correction orthographique.

Il y a peu d'études s'intéressant aux troubles orthographiques des dyslexiques par rapport aux études s'intéressant à leurs difficultés en lecture, même si l'on constate un accroissement du nombre de recherches sur l'orthographe des dyslexiques depuis quelques années (Cidrim & Madeiro, 2017). Pourtant les difficultés en orthographe des personnes dyslexiques persistent davantage que leurs difficultés en lecture. Dans une étude comparant les performances en orthographe d'étudiants francophones dyslexiques et non dyslexiques de même âge et de même niveau scolaire, (Mazur-Palandre, 2018) constate des profils d'erreurs similaires entre les deux groupes d'étudiants : les dyslexiques font les mêmes types d'erreurs que les non dyslexiques dans les mêmes proportions relatives. Cependant, les étudiants dyslexiques font significativement plus d'erreurs que les étudiants non dyslexiques. De plus, une analyse qualitative des erreurs révèle que les étudiants dyslexiques font des erreurs atypiques qui ne sont jamais retrouvées dans les écrits des étudiants non dyslexiques, notamment en ce qui concerne les accords et la conjugaison (ex : *les personnes proviennes*, *j'ai préférez*).

Les travaux sur la constitution et l'analyse de corpus écrits ont surtout pour objectif d'étudier l'apprentissage d'une langue étrangère (Granger, 2009) ou l'apprentissage de l'écrit dans la langue maternelle (Wolfarth *et al.*, 2016). L'exploration de corpus de productions écrites de dyslexiques pour le développement de correcteurs orthographiques concernent peu de travaux : (Pedler, 2007) pour l'anglais, (Rello *et al.*, 2012, 2014) pour l'espagnol, et (Antoine *et al.*, 2019) pour le français.

Dans un premier temps, nous présenterons un état de l'art de l'analyse des erreurs à partir de corpus écrits pour la correction orthographique. Puis nous décrirons les corpus de textes à notre disposition et la méthodologie utilisée pour leur analyse. Nous décrirons ensuite les résultats de l'analyse et les comparerons aux travaux déjà réalisés. Enfin, nous conclurons sur les implications des résultats sur le développement de modules de correction orthographique.

2 État de l'art

(Damerau, 1964) propose quatre type d'erreurs simples pour la correction orthographique de mots isolés (c'est-à-dire, sans prise en compte du contexte) : l'insertion d'un caractère, l'omission d'un caractère, la substitution d'un caractère par un autre, la transposition de deux caractères adjacents. Il trouve que plus de 80 % des mots qui ne sont pas dans un dictionnaire diffèrent du mot attendu d'une seule erreur de l'un de ces quatre types. Ces travaux ont abouti au développement de la distance d'édition de Damerau-Levenshtein permettant de calculer le nombre minimum d'opérations nécessaires pour transformer une chaine de caractères en une autre chaine et sont utilisés pour la correction orthographique de mots isolés.

Ce taux de 80 % d'erreurs à une distance d'édition de 1 de leur forme correcte n'est pas retrouvé par (Mitton, 1987) sur un corpus de productions d'élèves faibles en orthographe. Seulement 69 % des erreurs issues de ce corpus entrent dans les quatre catégories d'erreurs définies par Damerau. Les autres erreurs sont en majorité des errreurs non lexicales (*real-word error*), c'est-à-dire des erreurs qui aboutissent à un mot qui existe dans le dictionnaire et qui ne peuvent donc pas être détectées

et corrigées sans prise en compte du contexte environnant. Concernant les écrits des personnes dyslexiques, on peut s'attendre à ce que le taux d'erreurs à une distance d'édition supérieure à 1 de leur forme correcte et le taux d'erreurs non lexicales soient encore plus importants.

Quelques travaux se sont intéressés à l'analyse des erreurs produites par les personnes dyslexiques pour la correction orthographique.

(Pedler, 2007) a constitué un corpus de productions écrites de personnes dyslexiques en langue anglaise. Un premier échantillon de 3134 mots dont 636 sont erronés (20 % du corpus) a été analysé avec la typologie suivante :

- erreur simple : une seule opération d'édition est nécessaire (parmi les 4 opérations définies par Damerau) pour passer du mot erroné au mot attendu
- erreur multiple : plus d'une opération d'édition sont nécessaires pour passer du mot erroné au mot attendu
- erreur de segmentation : fusion (omission d'un espace) ou segmentation (insertion d'un espace)

Cet échantillon contient 53 % d'erreurs simples, 39 % d'erreurs multiples et 8 % d'erreurs de segmentation. Les erreurs non lexicales représentent 17 % des erreurs du corpus. Cette première analyse montre que les personnes dyslexiques anglophones font beaucoup d'erreurs multiples et que le nombre d'erreurs non lexicales qu'ils produisent n'est pas négligeable.

Par la suite, (Pedler, 2007) s'est intéressée en particulier aux erreurs non lexicales. Elle a constitué un second corpus rassemblant des documents d'origines diverses (devoirs à la maison d'élèves, rédactions d'étudiants, expérimentation de saisie de texte en ligne, forums de discussion et listes de diffusion sur Internet, etc.) pour développer et évaluer un correcteur orthographique dédié à la correction de ce type d'erreurs. Ce corpus contient 21 524 mots dont 2654 sont erronés. Les erreurs non lexicales représentent près d'un tiers des erreurs de ce second corpus. Pour détecter et corriger ce type d'erreurs, elle propose de construire une liste de plusieurs milliers d'ensembles de confusion (ensembles de mots souvent confondus comme *loose* et *lose*) combinée à une analyse syntaxique et sémantique pour déterminer quel mot dans l'ensemble de confusion est le plus probable dans le contexte. Cependant, elle exclut les erreurs d'accord et de conjugaison des ensembles de confusion et ne peut donc pas corriger toutes les erreurs non lexicales avec cette approche (plus d'un tiers des erreurs non détectées sont des erreurs d'accord et de conjugaison).

(Rello *et al.*, 2012) ont constitué le premier corpus de productions écrites de personnes dyslexiques en langue espagnole (castillan), DysCorpus. Ce corpus comprend 16 textes manuscrits écrits par des enfants dyslexiques de 6 à 15 ans. Il contient 1057 mots dont 157 sont erronés (15 % du corpus). En reprenant la méthodologie utilisée par (Pedler, 2007), ils trouvent 67 % d'erreurs simples, 23 % d'erreurs multiples et 10 % d'erreurs de segmentation. Les erreurs non lexicales représentent 21 % des erreurs du corpus. Les auteurs expliquent le plus faible taux d'erreurs multiples en espagnol par rapport à l'anglais par le fait que l'orthographe de l'espagnol est plus transparente que celle de l'anglais. Cependant, le taux d'erreurs non lexicales est similaire dans les deux langues. Cela confirme que ce type d'erreurs constitue un véritable problème pour la correction orthographique des écrits des personnes dyslexiques.

(Rello *et al.*, 2014) ont étendu ce corpus avec de nouveaux textes pour atteindre un corpus de 83 textes manuscrits également rédigés par des enfants dyslexiques de 6 à 15 ans. Ils ont extrait de ce nouveau corpus 887 mots erronés et 1171 erreurs dans une liste, DysList, qu'ils ont enrichi de nombreuses informations linguistiques : distance d'édition, fréquence, longueur, position de l'erreur, nombre de syllabes et structure syllabique, type d'erreur reprenant la typologie de Damerau, erreur lexicale

ou non lexicale, informations visuelles (ex : lettres miroirs), informations phonétiques (comme le voisement ou le point d'articulation des phonèmes), transfert linguistique chez les enfants bilingues (catalan/castillan). Les travaux réalisés sur ce corpus sont utilisés pour la création d'un correcteur orthographique (Rello *et al.*, 2015) pour les dys en langue espagnole. Appliqué à la détection et à la correction des erreurs non lexicales, ce correcteur détecte et corrige plus d'erreurs non lexicales que les correcteurs classiques, mais au prix d'une précision moindre (plus de faux positifs).

En langue française, (Antoine *et al.*, 2019) ont constitué un corpus de textes rédigés par 5 enfants dyslexiques et 5 enfants paralysés cérébraux pour un système d'aide à la communication combinant prédiction et correction orthographique. Ce corpus rassemble 521 erreurs orthographiques qui ont été annotées en suivant un schéma d'annotation répondant aux besoins des chercheurs en TAL et à ceux des orthophonistes. Pour la recherche en TAL, ils notent si le mot comporte une ou plusieurs erreurs distinctes, le type d'erreur (lexicale, syntaxique ou sémantique) et la morphologie en distinguant les erreurs de segmentation (fragmentation ou fusion) des autres erreurs pour lesquelles ils calculent la distance d'édition de Damerau-Levenshtein entre la forme erronée et la forme attendue. Pour les besoins des orthophonistes, ils établissent une typologie des erreurs en distinguant les erreurs phonologiquement plausibles (erreurs qui ne modifient pas la prononciation du mot, par exemple : *insi* (*ainsi*)) et les erreurs phonologiquement non plausibles (erreurs qui modifient la prononciation du mot, par exemple : *cantre* (*centre*)).

Leurs travaux, les premiers s'intéressant à la langue française, montrent un taux d'erreurs multiples de 54 % similaire à celui retrouvé par (Pedler, 2007) en anglais. Cependant, les taux d'erreurs non lexicales (29 %) et d'erreurs de segmentation (15 %) sont supérieurs à ceux retrouvés en anglais et en espagnol. Cela suggère que les problèmes rencontrés chez les dyslexiques anglais et espagnols sont également retrouvés, dans des proportions plus importantes, chez les dyslexiques français.

3 Méthodologie d'analyse des corpus

3.1 Description des corpus

Pour cette étude, nous avons utilisé deux corpus de productions écrites de personnes dyslexiques. Le premier corpus nous a été fourni par la FFDys [1], le second par une orthophoniste qui travaille avec des personnes dyslexiques en lien avec la FFDys.

Le premier corpus contient 9 textes scolaires (contrôles, exercices, dictées) écrits par des élèves dyslexiques de collèges et lycées (de la 5e à la terminale). Sept textes ont été écrits au clavier, les deux autres sont des textes manuscrits. Ils ont été écartés de la présente étude. En effet, le mode d'entrée du texte peut avoir un impact sur le type d'erreurs produites. (Sitbon *et al.*, 2007) constatent que certaines erreurs rencontrées dans les textes manuscrits d'enfants dyslexiques, telles que les substitutions de lettres miroirs (p/q ou b/d par exemple), ne sont pas observées sur des textes écrits au clavier. De plus, l'utilisation du clavier entraine des erreurs de frappe qu'on ne retrouvera pas dans les textes manuscrits. Les 7 textes écrits au clavier totalisent 3357 mots [2]. Ce sont des textes relativement longs (475 mots en moyenne par texte). Ce premier corpus contient 1240 formes erronées [3] dont 771 formes

1. Fédération Française des Dys
2. Nous entendons par mot toute séquence de caractères séparée par des espaces ou de la ponctuation.
3. Nous préférons parler de formes erronées plutôt que de mots erronés. Une forme erronée peut correspondre à un ou plusieurs mots. Ex : *plus par* (*plupart*)

distinctes.

Le second corpus contient 71 textes courts (53 mots en moyenne par texte) écrits au clavier par des personnes dyslexiques âgées de 16 à 45 ans (âge moyen = 22,5 ans, écart-type = 4,7 ans). Ce corpus est lui-même composé de :

— 6 dictées
— 33 expressions écrites dirigées
— 32 expressions écrites libres

Il totalise 3913 mots et 879 formes erronées dont 594 formes distinctes.

L'ensemble des deux corpus totalisent 7270 mots et 2119 formes erronées dont 1303 distinctes.

3.2 Annotation des erreurs

Pour chaque texte, nous avons extrait manuellement les formes erronées dans un tableau, puis pour chaque forme erronée, nous avons noté :

— la forme erronée
— la forme attendue
— le lemme [4] de la forme attendue
— la phrase contenant la forme erronée
— le nombre d'erreurs et leurs types
— la distance d'édition de Damerau-Levenshtein entre la forme erronée et la forme attendue
— la similarité entre les transcriptions phonétiques des formes erronée et attendue
— si l'erreur est lexicale ou non-lexicale
— le nombre de mots erronés dans le contexte (les 2 mots précédents et les 2 mots suivants)

3.3 Les différents types d'erreurs

Au lieu de distinguer comme (Pedler, 2007) et (Rello *et al.*, 2012) les erreurs simples et les erreurs multiples, nous calculons, d'une part, la distance d'édition entre la forme erronée et la forme attendue et, d'autre part, nous comptons le nombre d'erreurs de la forme erronée comme (Antoine *et al.*, 2019).

Pour les types d'erreurs, nous utilisons la typologie présentée dans la table 1. Cette typologie s'inspire de celle définie par (Plisson & Daigle, 2013) pour décrire les erreurs d'enfants dyslexiques francophones. Par rapport à cette typologie, nous ne distinguons pas les erreurs phonologiquement plausibles des erreurs non phonologiquement plausibles. Nous regroupons dans une même catégorie les erreurs de phonétisation concernant les mauvais choix de graphèmes [5], les lettres muettes et les morphogrammes lexicaux [6]. L'idée étant que ces différents types d'erreurs peuvent être corrigés avec la même approche. De même, nous comptabilisons les erreurs concernant les traits d'union dans les erreurs de segmentation plutôt qu'avec les erreurs sur les majuscules.

4. Le lemme d'un mot est sa forme canonique telle qu'on la trouve dans un lexique. Préciser le lemme permet de distinguer les homographes tels que *est* forme conjuguée du lemme *être* et *est* point cardinal).

5. (Catach, 1986) définit le graphème comme la plus petite unité distinctive de la chaine écrite et le phonème comme la plus petite unité distinctive de la chaine orale. Par exemple, le mot *châteaux* se décompose en 5 graphèmes : ‹ ch ›, ‹ â ›, ‹ t ›, ‹ eau › et ‹ x › et en 4 phonèmes : /ʃ/, /ɑ/, /t/ et /o/.

6. (Catach, 1986) définit le morphogramme lexical comme un graphème non chargé de transcrire un phonème et permettant d'établir un lien avec les dérivés. Par exemple, le ‹ t › final dans *petit*.

En fonction de son type, une erreur peut concerner l'ensemble d'un mot (ex : confusion entre les homophones *ces* et *c'est*), un graphème (ex : substitution du graphème ‹ss› par le graphème ‹s› dans *réusite*) ou un caractère (ex : omission de l'apostrophe dans *lafrique*).

Type d'erreurs	**Exemples**
Phonétisation : mauvais choix de graphème et lettre muette	comerse (commerce), toujour (toujours)
Substitution d'un graphème par un autre phonétiquement proche	réusite (réussite)
Confusion entre homophones	ces (c'est)
Erreur d'accord en genre et nombre et de conjugaison	autre (autres), rajouterai (rajouterait)
Erreur de segmentation : fragmentation ou fusion (incluant les erreurs concernant les apostrophes et les traits d'union)	quel que (quelque), ducou (du coup), lafrique (l'Afrique), rendévous (rendez-vous)
Liaison erronée	on na (on a)
Majuscule	japon (Japon)
Ajout d'un caractère	situiation (situation)
Omission d'un caractère	Qustion (Question)
Substitution d'un caractère par un autre caractère	dont (sont)
Transposition de deux caractères adjacents	aprle (parle)
Déplacement d'un caractère	disgetif (digestif)
Omission ou répétition de mot	il trouve pas (il ne trouve pas)
Mauvais choix lexical	famille (familiale)
Mot non reconnu	sanéte

TABLE 1: Types d'erreurs

4 Résultats

4.1 Distance de Damerau-Levenshtein

La table 2 présente les pourcentages de formes erronées à une distance de 1, 2 ou plus de leur forme correcte pour chaque corpus et pour les deux corpus. En moyenne sur les deux corpus, une large proportion de formes erronées (41 %) sont à une distance de 2 ou plus de leur forme correcte. On note cependant une différence importante entre les deux corpus : sur le premier corpus, un peu moins de la moitié des formes sont concernées, un tiers des formes sur le second corpus.

Dans le premier corpus, la distance maximum est de 7 et concerne deux formes erronées : *oré* (*auraient*) et *nalé* (*n'allaient*). Dans le deuxième corpus, la distance maximum est de 5 et concerne trois formes erronées : *fesé* (*faisais*), *noyer* (*nettoyé*) et *setoufle* (*s'étouffent*).

Corpus	Distance = 1	Distance = 2	Distance > 2
1	53 %	24,7 %	22,3 %
2	67,1 %	21,8 %	11,1 %
1 et 2	58,8 %	23,5 %	17,7 %

TABLE 2: Distance de Damerau-Levenshtein

4.2 Similarité phonétique

Dans un premier temps, nous avons comparé la transcription phonétique des formes erronées et attendues. Ces transcriptions ont été obtenues grâce au transcripteur LIA_PHON (Béchet, 2001). Puis, nous avons comparé les phonétiques après simplification de la phonétique des voyelles. Nous avons réduit le nombre de voyelles prises en compte par LIA_PHON de 15 à 10 : nous ne distinguons plus les voyelles /e/ et /ɛ/ (dans *thé* et *cette*), /o/ et /ɔ/ (dans *tôt* et *botte*), /ø/, /œ/ et /ə/ (dans *peu*, *peur* et *le*), et /ɛ̃/ et /œ̃/ (dans *brin* et *brun*). En effet, suivant la personne ou la région, la prononciation des voyelles peut varier (par exemple, *très* est prononcé [tʁe] ou [tʁɛ]) et certaines oppositions peuvent disparaitre (pas de distinction entre *brin* et *brun* par exemple).

La table 3 présente les pourcentages de formes erronées ayant la même phonétique que leur forme correcte (c'est-à-dire, les erreurs phonologiquement plausibles) pour chaque corpus et pour les deux corpus. Dans les deux corpus, plus de la moitié des formes erronées ont une phonétique identique à celle de leur forme correcte. Si on utilise la phonétique simplifiée, deux tiers des formes erronées ont une phonétique proche de celle de leur forme correcte.

Corpus	Phonétique	Phonétique simplifiée
1	58,9 %	69,7 %
2	58,5 %	63,4 %
1 et 2	58,7 %	67,1 %

TABLE 3: Similarité phonétique

4.3 Erreurs non lexicales

Une erreur non lexicale est une erreur qui produit un mot présent dans le lexique. Il s'agit essentiellement d'erreurs syntaxiques (ex : *les région* au lieu de *les régions*) et sémantiques (ex : *famille* au lieu de *familial*). Plus rarement, les erreurs de segmentation peuvent produire des erreurs non lexicales (ex : *plus par* au lieu de *plupart*, *lest* au lieu de *l'est*).

Le choix du lexique qui sert à la correction orthographique est important. Plus celui-ci est large plus il va contenir des formes rares, peu usitées et plus le risque qu'une forme erronée se retrouve dans le lexique augmente. Par exemple : Les formes erronées *oré* (*auraient*), *mayeur* (*meilleur*) et *este* (*Est*) sont dans le lexique Morphalou 3. Pour détecter l'erreur, il faut alors utiliser une analyse syntaxique voire sémantique.

Nous avons comparé 3 lexiques :

— Morphalou (version 3.1) (ATILF, 2019) : un lexique à large couverture qui aggrège plusieurs lexiques pour atteindre 954 690 formes fléchies

— Dicollecte (version 6.4.1)[7] : un lexique de plus de 500 000 formes fléchies utilisé par le correcteur orthographique Hunspell en français
— Lexique (version 3.83) (New *et al.*, 2004) : un lexique de plus de 140 000 formes fléchies

La table 4 présentent les pourcentages d'erreurs non lexicales relevées dans les corpus en fonction du lexique choisi. Quel que soit le lexique utilisé, sur l'ensemble des deux corpus, un peu plus de la moitié des formes erronées sont des erreurs non lexicales. Même si le deuxième corpus contient une proportion plus faibles d'erreurs que le premier corpus, le pourcentage d'erreurs non lexicales y est plus élevé.

Corpus	Morphalou 3	Dicollecte	Lexique 3
1	607 (49 %)	595 (48 %)	587 (48 %)
2	531 (60 %)	523 (59 %)	523 (59 %)
1 et 2	1138 (54 %)	1118 (53 %)	1110 (53 %)

TABLE 4: Erreurs non lexicales

4.4 Formes correctes les plus souvent erronées

Les 10 formes correctes les plus fréquemment erronées sont des mots courts (1 à 5 lettres), le plus souvent monosyllabiques (à l'exception de *après* qui est constitué de 2 syllabes, ils possèdent tous 1 seule syllabe) et fréquents. La table 5 présente les 10 formes correctes les plus fréquemment erronées, le nombre d'occurrences erronées, le pourcentage de formes erronées et les différentes formes erronées.

Forme correcte	Nombre d'occurrences erronées	Pourcentage d'occurences erronées	Formes erronées
très	21	87,5 %	tré, tres
peut	13	86,7 %	pue, peu, pela
à	115	81,6 %	a, d, ∅
après	12	80 %	apres, apré, apra, apre, a prais, apret
ils	12	73,3 %	il
ont	13	72,2 %	on
c'est	21	58,3 %	ses, sé, ces, s'est, cces
ce	18	42,9 %	se, si
au	15	29,4 %	o, a
est	23	28,4 %	et, é, n'ait, ai, soi, ∅

TABLE 5: Formes les plus fréquemment erronées

4.5 Erreurs sur la première lettre

D'après (Yannakoudakis & Fawthrop, 1983), la première lettre d'un mot erroné est correcte dans la majorité des cas en anglais (moins de 2 % des erreurs sont retrouvées à l'initiale des mots).

7. Ce lexique est téléchargeable à l'adresse https ://grammalecte.net/download.php ?prj=fr

Dans l'ensemble de nos corpus, nous avons 16,5 % de formes dont la première lettre est erronée (10,9 % si on exclut les mots d'une seule lettre). On ne prend pas en compte les erreurs de majuscule.

Si l'on regarde la phonétique, moins de 4 % des formes erronées sont phonétiquement incorrectes à l'initiale.

4.6 Variabilité des formes erronées

Dans l'ensemble du corpus, 200 formes correctes ont au moins deux formes erronées (18,3 % des 1092 formes correctes distinctes).

On compte jusqu'à 6 formes erronées différentes dans l'ensemble du corpus pour la forme *après*.

4.7 Contexte autour du mot

La correction orthographique nécessite souvent une analyse contextuelle :

- pour les erreurs lexicales : pour sélectionner la meilleure correction parmi une liste de corrections potentielles
- pour les erreurs non lexicales : pour les détecter et les corriger

Cependant, si le contexte autour du mot est erroné, l'analyse contextuelle peut donner des résultats erronés.

Pour chaque mot erroné nous avons regardé, si le contexte local (2 mots avant et 2 mots après) était correct ou erroné. La table 6 présente les proportions de formes erronées avec aucun, un ou plusieurs mots erronés dans leur contexte. 72,3 % des formes erronées ont au moins un mot de contexte erroné.

Nombre de mots erronés	Pourcentage de formes erronées concernées
0	27,7 %
1	39,5 %
2	24,4 %
3	7,4 %
4	1,0 %

TABLE 6: Nombre de mots de contexte erroné

4.8 Répartitions des erreurs dans les différentes catégories

Le nombre moyen d'erreurs par mot est de 1,4 sur l'ensemble des corpus. Le premier corpus a en moyenne 1,48 erreurs par mot (écart-type = 0,13), le second corpus a en moyenne 1,24 erreurs par mot (écart-type = 0,19). Le nombre maximum d'erreurs par mot est de 5 pour le premier corpus et de 4 pour le second corpus. Une forme erronée peut combiner plusieurs erreurs de types différents. Par exemple, la forme erronée *meiu* (*mieux*) combine une transposition de caractères adjacents et une omission de lettre muette.

La répartition des erreurs en fonction de leur type est présentée dans la table 7. Les erreurs les plus fréquentes sont les erreurs de phonétisation et les erreurs d'accord en genre et nombre et de conjugaison. Ces deux catégories d'erreurs représentent plus de la moitié des erreurs.

Type d'erreurs	Exemples	Pourcentage d'erreurs
Phonétisation : mauvais choix de graphème et lettre muette	comerse (commerce), toujour (toujours)	27,25 %
Erreur d'accord en genre et nombre et de conjugaison	autre (autres), rajouterai (rajouterait)	26,81 %
Substitution d'un graphème par un autre phonétiquement proche	réusite (réussite)	15,98 %
Confusion entre homophones	ces (c'est)	11,28 %
Erreur de segmentation : fragmentation ou fusion (incluant les erreurs concernant les apostrophes et les traits d'union)	quel que (quelque), ducou (du coup), lafrique (l'Afrique), rendévous (rendez-vous)	6,35 %
Majuscule	japon (Japon)	3,14 %
Omission d'un caractère	Qustion (Question)	3,04 %
Substitution d'un caractère par un autre caractère	dont (sont)	1,59 %
Omission ou ajout de mot	il trouve pas (il ne trouve pas)	1,28 %
Ajout d'un caractère	situiation (situation)	1,25 %
Transposition de deux caractères adjacents	aprle (parle)	1,01 %
Mauvais choix lexical	famille (familiale)	0,54 %
Mot non reconnu	sanéte	0,17 %
Liaison erronée	on na (on a)	0,20 %
Déplacement d'un caractère	disgetif (digestif)	0,07 %

TABLE 7: Répartition des erreurs

5 Discussion des résultats

Les corpus que nous avons analysés contiennent environ un tiers de mots erronés. Ce taux est plus important que ceux trouvés dans les travaux de (Pedler, 2007) en anglais (20 % de mots erronés) et de (Rello *et al.*, 2012) en espagnol (15 % de mots erronés). Il est cependant plus faible que celui trouvé par (Antoine *et al.*, 2019) en français (55 % de mots erronés). Nous observons également un nombre moyen d'erreurs par mot inférieur à celui observé par (Antoine *et al.*, 2019) : 1,4 contre 1,8 erreurs par mot.

Concernant la distance d'édition, nous avons un taux d'erreurs multiples de 41,2 %, supérieur aux 23 % de (Rello *et al.*, 2012), mais proche des 39 % de (Pedler, 2007). De nouveau, (Antoine *et al.*, 2019) trouve un taux supérieur (54 %).

Nous avons également un taux d'erreurs de segmentation proche de celui observé par (Pedler, 2007) (respectivement 6 % et 8 %). (Rello *et al.*, 2012) et (Antoine *et al.*, 2019) trouvent des taux légèrement plus élevés (respectivement 10 % et 15 %). Même si le taux d'erreurs de segmentation que nous relevons est relativement faible, ces erreurs ne doivent pas être négligées car elles peuvent être particulièrement difficiles à corriger quand elles produisent des erreurs non lexicales (*laide* au lieu de

l'aide) ou qu'elles sont cumulées à d'autres erreurs (*apeures* au lieu de *à peu près*).

Les taux plus faibles que nous observons par rapport aux résultats de (Antoine *et al.*, 2019) peuvent s'expliquer par le fait que nous avons des textes rédigés par des adultes ou des élèves de collège et lycée alors que l'âge moyen des enfants dyslexiques de l'étude de (Antoine *et al.*, 2019) est de 10 ans. Nous observons également une différence entre notre premier corpus constitué d'écrits de collégiens et de lycéens et notre second corpus constitués d'écrits rédigés par des personnes plus âgées : les textes du premier corpus contiennent en moyenne plus de formes erronées, plus d'erreurs par mot et plus d'erreurs multiples. Cela pose la question de savoir quelles sont les compétences en orthographe qui peuvent être améliorées par l'apprentissage chez les personnes dyslexiques.

Nous observons un taux d'erreurs non lexicales nettement supérieur aux autres langues : plus de la moitié des formes erronées de nos corpus sont présentes dans les trois lexiques que nous avons testés, contre 17 % pour (Pedler, 2007) en anglais et 9 % pour (Rello *et al.*, 2014) en espagnol. Une analyse plus fine des erreurs non lexicales de nos corpus est nécessaire pour comprendre cette différence avec les autres langues, notamment pour connaitre la proportion d'erreurs syntaxiques et d'erreurs sémantiques parmi ces erreurs non lexicales. À titre d'exemple, (Antoine *et al.*, 2019) trouvent 29 % d'erreurs non lexicales dans leur corpus en français.

Étant donné le nombre très élevé d'erreurs non lexicales, une analyse contextuelle est indispensable. L'analyse de notre corpus montre que le contexte contient souvent des mots erronés ce qui peut perturber l'analyse contextuelle. Nous nous sommes pour l'instant contenter de regarder si le contexte local (les 2 mots précédents et les 2 mots suivants) était erroné. L'analyse du contexte peut être poussée plus loin en regardant la proportion d'erreurs pouvant être corrigées dans un contexte local et celles nécessitant une analyse plus globale comme la phrase, en examinant les performances des étiqueteurs morpho-syntaxiques sur des phrases issues de nos corpus ou encore en regardant si le contexte est phonétiquement correct.

Nous trouvons un taux d'erreurs sur la première lettre de 16 % similaire à celui retrouvé par (Antoine *et al.*, 2019) (14 %) et légèrement supérieur à ceux observés par (Rello *et al.*, 2012) et (Pedler, 2007) (respectivement 10 % et 5 %). Ce taux reste assez faible. Cela ne confirme donc pas l'hypothèse avancée par (Bacquelé, 2015) selon laquelle les dyslexiques auraient des difficultés pour écrire correctement les initiales des mots. De plus, il est possible de s'appuyer sur la phonétique qui est correcte sur la première lettre dans plus de 96 % des cas.

Le taux de formes erronées phonétiquement similaires à leur forme correcte est de 59 %. Ce résultat est conforme aux observations de (Antoine *et al.*, 2019) qui trouvent 62 % d'erreurs phonologiquement plausibles. Cela suggère que le passage à la phonétique peut être intéressant pour corriger certaines erreurs comme le proposent (Sitbon *et al.*, 2007). En simplifiant les règles de transcription en phonétique des voyelles, nous trouvons deux tiers d'erreurs phonétiquement similaires à leur forme erronée. Il faut cependant trouver un juste milieu : plus on simplifie les règles de transcription en phonétique, plus on augmente le nombre de mots du lexique phonétiquement similaires aux mots erronés.

Concernant la variabilité des formes erronées, notre analyse montre une variabilité inter-individuelle parfois importante. Cependant, nous n'avons pas pu analyser la variabilité intra-individuelle car nous ne disposons pas d'assez de textes écrits par une même personne. Cette question est toutefois intéressante : si une même personne dyslexique reproduit souvent le même type d'erreurs, il est peut-être possible de modéliser son profil d'erreurs.

Enfin, il serait également intéressant d'analyser plus précisément les difficultés communes à toutes

les personnes dyslexiques quelle que soit la langue et celles qui sont spécifiques à chaque langue. Par exemple, le français est une langue à l'orthographe plutôt opaque : les relations entres phonèmes et graphèmes sont irrégulières. En particulier, dans le sens de l'écriture, le français est proche de l'anglais, langue à l'orthographe très opaque. Une étude comparant les erreurs produites par des dyslexiques grecs et des dyslexiques américains suggère que les différentes caractéristiques d'une langue entrainent différents types et proportions d'erreurs : les dyslexiques grecs (dont la langue est considérée comme transparente) font en proportion significativement moins d'erreurs phonologiquement non plausibles que les dyslexiques américains, mais plus d'erreurs phonologiquement plausibles et plus d'erreurs grammaticales (Giannouli & Pavlidis, 2014). Une autre particularité du français est le décalage important entre la morphologie de l'oral et celle de l'écrit (Jaffré, 2005) : les marques de genre et de nombre, les terminaisons verbales sont très présentes à l'écrit, mais quasiment absentes à l'oral. Ce n'est pas le cas de l'anglais ou de l'espagnol et cela peut constituer une difficulté supplémentaire pour les dyslexiques francophones.

6 Conclusion et travaux futurs

Dans cet article, nous avons présenté une analyse des erreurs produites par des personnes dyslexiques françaises (collégiens, lycéens et adultes) à partir de corpus écrits. Comme (Antoine *et al.*, 2019), nous retrouvons des taux de mots erronés, d'erreurs multiples et d'erreurs non lexicales supérieurs à ceux observés dans des corpus de personnes dyslexiques anglaises ou espagnoles.

Cette étude soulève plusieurs points importants :

— le nombre très élevé d'erreurs non lexicales : une analyse contextuelle est indispensable pour détecter et corriger efficacement ces erreurs
— le nombre d'erreurs phonétiquement similaires ou proches de leur forme correcte étant important, dans quelle mesure peut-on s'appuyer sur la phonétique pour corriger les erreurs ?
— le contexte pouvant lui-même contenir des erreurs, dans quelle mesure peut-on s'appuyer sur le contexte pour détecter et corriger les erreurs ?

La prochaine étape consistera à évaluer les performances des correcteurs orthographiques actuels sur ce corpus pour identifier les erreurs qu'ils parviennent à corriger et celles qu'ils ne parviennent pas à détecter ou corriger. Nous disposons actuellement d'un premier module de correction basique fondé sur la phonétique et nous comparerons les résultats obtenus par notre module avec ceux des correcteurs.

Références

ANTOINE J.-Y., CROCHETET M., ARBIZU C., LOPEZ E., POUPLIN S., BESNIER A. & THEBAUD M. (2019). Ma copie adore le vélo : analyse des besoins réels en correction orthographique sur un corpus de dictées d'enfants. In *TALN 2019*, Toulouse, France. HAL : hal-02375246.

ATILF (2019). Morphalou. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

BACQUELÉ V. (2015). L'usage de l'informatique par les élèves dyslexiques : un outil de compensation à l'épreuve de l'inclusion scolaire. *Terminal. Technologie de l'information, culture & société*, (116). DOI : 10.4000/terminal.661.

BARROUILLET P., BILLARD C., AGOSTINI M. D., DÉMONET J.-F., FAYOL M., GOMBERT J.-E., HABIB M., NORMAND M.-T. L., RAMUS F., SPRENGER-CHAROLLES L. & VALDOIS S. (2007). *Dyslexie, dysorthographie, dyscalculie : bilan des données scientifiques*. Rapport de recherche, INSERM. HAL : hal-01570674.

BÉCHET F. (2001). LIA_phon : un système complet de phonétisation de textes. *Traitement Automatique des Langues*, **42**(1), 47–67.

CATACH N. (1986). *L'orthographe française : traité théorique et pratique : avec des travaux d'application et leurs corrigés (avec la collab. de Claude Gruaz et Daniel Duprez)*. Paris, nathan édition.

CIDRIM L. & MADEIRO F. (2017). Studies on spelling in the context of dyslexia : a literature review. *Revista CEFAC*, **19**(6), 842–854. DOI : 10.1590/1982-0216201719610317.

DAMERAU F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, **7**(3), 171–176. DOI : 10.1145/363958.363994.

GIANNOULI V. & PAVLIDIS G. T. (2014). What can spelling errors tell us about the causes and treatment of dyslexia ? : What can Spelling Errors Tell Us about the Causes and Treatment of Dyslexia ? *Support for Learning*, **29**(3), 244–260. DOI : 10.1111/1467-9604.12065.

GRANGER S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching. *Corpora and language teaching*, **33**, 13–32. DOI : 10.1075/scl.33.04gra.

JAFFRÉ J.-P. (2005). L'orthographe du français, une exception ? *Le francais aujourd'hui*, **n° 148**(1), 23–31. DOI : 10.3917/lfa.148.0023.

MAZUR-PALANDRE A. (2018). La dyslexie à l'âge adulte : la persistance des difficultés orthographiques. *SHS Web of Conferences*, **46**, 10003. DOI : 10.1051/shsconf/20184610003.

MITTON R. (1987). Spelling checkers, spelling correctors and the misspellings of poor spellers. *Information Processing & Management*, **23**(5), 495–505. DOI : 10.1016/0306-4573(87)90116-6.

NEW B., PALLIER C., BRYSBAERT M. & FERRAND L. (2004). Lexique 2 : A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, **36**(3), 516–524. DOI : 10.3758/BF03195598.

PEDLER J. (2007). *Computer Correction of Real-word Spelling Errors in Dyslexic Text*. Thèse de doctorat, University of London.

PLISSON A. & DAIGLE, DANIEL ANDYA MONTESINOS-GELET I. (2013). The Spelling Skills of French-Speaking Dyslexic Children. *Dyslexia*, **19**(2), 76–91. DOI : 10.1002/dys.1454.

RELLO L., BAEZA-YATES R. & LLISTERRI J. (2014). DysList : An Annotated Resource of Dyslexic Errors. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, p. 1289–1296, Reykjavik, Iceland : European Language Resources Association (ELRA). DOI : 10.13140/2.1.2542.7205.

RELLO L., BAEZA-YATES R., SAGGION H. & PEDLER J. (2012). A First Approach to the Creation of a Spanish Corpus of Dyslexic Texts. In *LREC Workshop Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, p. 22–27, Turkey.

RELLO L., BALLESTEROS M. & BIGHAM J. P. (2015). A Spellchecker for Dyslexia. p. 39–47 : ACM. DOI : 10.1145/2700648.2809850.

SITBON L., BELLOT P. & BLACHE P. (2007). Traitements phrastiques phonétiques pour la réécriture de phrases dysorthographiées. In *TALN 2007*, Toulouse, France. HAL : hal-01321119.

WOLFARTH C., PONTON C. & BRISSAUD C. (2016). Du TAL dans les écrits scolaires : premières approches. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016*, volume 09, p. 30–37, Paris. HAL : hal-01878718.

YANNAKOUDAKIS E. J. & FAWTHROP D. (1983). The rules of spelling errors. *Information Processing & Management*, **19**(2), 87–99. DOI : 10.1016/0306-4573(83)90045-6.