Un comparateur phonétique dissymétrique pour la correction orthographique à destination des dyslexiques

Johana Bodard

Un comparateur phonétique dissymétrique pour la correction orthographique à destination des dyslexiques

Johana Bodard

CHArt-THIM (EA 4004), Université Paris 8, Saint-Denis, France
johana.bodard@etud.univ-paris8.fr

Résumé

La correction orthographique de textes écrits par des personnes dyslexiques-dysorthographiques (dys) constitue un défi. En effet, les spécificités des fautes produites, le nombre important de fautes par mots et la variabilité intra-individuelle mettent en échec les correcteurs orthographiques classiques. Cet article présente une méthode pour la correction orthographique chez les dys. Cette méthode se fonde sur une comparaison des mots saisis à ceux d'un dictionnaire après les avoir préalablement convertis en phonétique. La conversion en phonétique des mots saisis diffère de celle des mots du dictionnaire par l'intégration des confusions phonétiques fréquentes chez les dys. En prospective, l'impact positif sur la récupération des mots pousse à l'approfondissement et au complément par des méthodes prenant en compte les mots autour du mot saisi.

1 Introduction

La dyslexie et la dysorthographie sont des troubles cognitifs qui entrainent des difficultés importantes en lecture et en écriture. Ces deux troubles sont quasiment constamment associés [9]. Les difficultés apparaissent dans l'enfance lors des premiers apprentissages et persistent à l'âge adulte. En France, la prévalence est estimée entre 6 % et 8 % [3].

L'une des principales difficultés des personnes dyslexiques-dysorthographiques (dys) réside dans l'orthographe. Les personnes dys rencontrent notamment des difficultés dans l'orthographe des mots irréguliers ou nouveaux et dans l'application des règles permettant de convertir les graphèmes (les plus petites unités distinctives de la chaine écrite) en phonèmes (les plus petites unités distinctives de la chaine orale) et inversement. En effet, un phonème peut être retranscrit par plusieurs graphèmes (par exemple, en français, le phonème /o/ peut être retranscrit par les graphèmes *o*, *au*, *eau*, *op*, etc.), et un graphème peut correspondre à plusieurs phonèmes (le graphème *ch* peut se prononcer /ʃ/ dans *cheval* ou /k/ dans *psychologie*).

1.1 Nature des fautes commises

On peut identifier certaines fautes récurrentes à travers les situations d'écriture par des personnes dys :

— confusion entre phonèmes proches : « il fautra » (il faudra)
— omission des lettres muettes : « boneur » (bonheur), « soiré » (soirée)
— écriture fortement phonétisée : « des baignès » (des beignets), « je fesé » (je faisais)
— mauvaise segmentation des mots : « jevideo » (jeux vidéo), « la plus part » (la plupart)
— ajout, suppression, inversion ou substitution de lettres : « un entretient » (un entretien), « il prnd » (il prend), « setp » (sept), « bousse » (douce)
— non respect des accords en genre, nombre, et de conjugaison : « une personne sportif » (une personne sportive), « ils chantais » (ils chantaient)
— apostrophe manquante ou en trop : « à laide » (à l'aide), « l'argement » (largement)

À la suite de contacts avec un réseau d'orthophonistes, nous avons pu disposer de corpus de textes rédigés par des personnes dys (collégiens, lycéens ou adultes), présélectionnés par les orthophonistes pour la variabilité dans la nature des fautes. Chaque texte correspond à un seul individu. Sur la base des 34 premiers textes (1484 mots), nous avons constaté que plus de la moitié des mots erronés contiennent plus de deux fautes et qu'un mot sur trois en moyenne est erroné. On constate en outre une forte variabilité intra-individuelle : un même mot va être écrit différemment d'une phrase à l'autre, d'un moment à l'autre [3] [13].

1.2 Les correcteurs orthographiques

Les correcteurs orthographiques font partie des outils technologiques recommandés aux personnes dys pour compenser leurs difficultés. Cependant, ces correcteurs ne sont pas adaptés à un usage spécifique par des personnes souffrant de troubles dys [2]. En effet, les algorithmes utilisés par les logiciels de correction classique sont développés avec l'hypothèse que :

— la source ne témoigne pas d'une intention de non-respect de l'orthographe, de la grammaire et de la syntaxe ;
— les fautes portent soit sur un nombre peu élevé de caractères par mot (fautes de frappe, oublis, erreurs d'OCR, hésitations telles qu'application d'une simple ou double consonne, etc.), soit sur l'accord en genre et en nombre ou la conjugaison.

Ils sont donc facilement mis en échec par les erreurs spécifiques des dys (écriture phonétique, erreurs de segmentation, nombre important d'erreurs par mots, etc.). Ils ignorent les mots erronés comme des mots inconnus ou génèrent des propositions erronées qui ne correspondent pas à ce que la personne souhaitait écrire [14]. Pour la personne dys, la tâche de correction devient alors rapidement fastidieuse, voire impossible. Des algorithmes spécifiques doivent donc être développés et expérimentés.

Nous présentons ici une première approche pour la correction orthographique dans un contexte de dyslexie-dysorthographie. Cette approche se fonde sur une comparaison des mots saisis à ceux d'un dictionnaire après les avoir préalablement convertis en phonétique. Nous proposons de prendre en compte, dans la phase de conversion en phonétique des mots saisis, les erreurs de conversion graphèmes-phonèmes et d'approximation phonétique les plus couramment produites par les dys (confusions entre paires de consonnes ou de voyelles dont la prononciation est proche).

2 Correction orthographique de mots isolés

Nous nous intéressons ici uniquement à la correction orthographique de mots isolés, c'est-à-dire sans prise en compte du contexte. La correction de mots isolés est en effet une première étape permettant d'obtenir pour un mot erroné une liste de corrections potentielles qui pourront par la suite être départagées à l'aide du contexte du mot (mots précédents, mots suivants, phrase contenant le mot à corriger).

2.1 Utilisation de mesures de similarité

L'approche la plus simple consiste à comparer chaque mot du texte saisi à ceux d'un dictionnaire et à utiliser une mesure de similarité pour calculer un indice de ressemblance entre un mot inconnu et les mots du dictionnaire. Les mots du dictionnaire qui sont les plus proches du mot inconnu sont les corrections potentielles. Levenshtein [12] propose un algorithme qui

permet de calculer la distance entre deux chaines de caractère en comptant le nombre d'opérations nécessaires pour passer d'une chaine à une autre (les opérations possibles étant l'ajout, la suppression ou le remplacement d'un caractère par un autre). Damerau propose un algorithme similaire en ajoutant l'opération de transposition de deux caractères adjacents aux opérations d'édition possibles [7]. Pour limiter le nombre de corrections potentielles, sont généralement conservés uniquement les mots à une distance de 1 du mot erroné (maximum 2 pour les mots longs). Par exemple, le correcteur orthographique Ispell propose uniquement les mots qui sont à une distance de Damerau-Levenshtein de 1 du mot erroné [11]. Parmi de nombreux algorithmes d'approche informatique ou linguistique, cet algorithme est très utilisé par les correcteurs orthographiques courants qui supposent que le nombre d'erreurs par mots est faible voire le plus souvent unique.

Dans le cas de la dyslexie-dysorthographie, les mots erronés sont parfois très éloignés de leur forme correcte et une simple application de ces algorithmes s'avère inefficace. Par exemple, dans le cas de *faisais* incorrectement orthographié *fesé*, l'application de la distance de Levenshtein avec une distance maximum de 1 renvoie les suggestions suivantes : *pesé*, *fusé*, *fessé*. La distance de Levensthein avec la forme correcte *faisais* est égale à 5, alors que phonétiquement les deux formes sont très proches. Cela suggère que pour corriger ce type de fautes, très fréquentes chez les dys, il faut utiliser une approche qui passe par une transcription phonétique des mots saisis.

2.2 Transcription phonétique

La transcription phonétique est une technique permettant de convertir un mot ou un texte en une version phonétique plus ou moins approximative selon l'approche utilisée.

Techniques fondées sur la similarité phonétique

Ces techniques permettent d'indexer des mots en fonction de leur prononciation. Le but est que deux mots à la prononciation identique ou proche soit encodés avec le même index même si leur orthographe diffère. Russel et Odell sont les premiers à proposer une telle technique avec l'algorithme phonétique Soundex pour indexer des noms propres en fonction de leur prononciation en anglais américain [17]. Par la suite, d'autres algorithmes reposant sur le même principe sont développés tels que Metaphone [15], qui prend en compte davantage de particularités de la prononciation anglaise et permet un encodage plus précis que Soundex, et Double Metaphone [16], une version de Metaphone qui introduit certaines spécificités d'autres langues que l'anglais comme le français.

Ces algorithmes phonétiques sont utilisés pour la recherche tolérante aux erreurs dans des bases de données ou pour la correction orthographique. Par exemple, le correcteur orthographique Aspell utilise une version modifiée de Double Metaphone [1].

Cependant, ces méthodes sont par nature approximatives. En particulier, les voyelles à l'intérieur des mots sont souvent simplement ignorées. Ces algorithmes commencent à être pertinents avec des mots suffisamment longs (5 consonnes ou plus). Ainsi, si *fesé* et *faisais* ont bien le même code Double Metaphone (*FS*), c'est aussi le cas de *fuseau*, *fausse*, *voisée*, etc.

Techniques fondées sur une transcription graphèmes-phonèmes

À l'opposé des algorithmes phonétiques présentés ci-dessus, les techniques de transcription graphèmes-phonèmes permettent de convertir un texte en phonétique en étant le plus proche de la prononciation réelle. Elles sont utilisées pour obtenir des représentations phonémiques de textes, exploitables dans des applications de synthèse vocale ou de reconnaissance de la parole.

Elles utilisent généralement une approche fondée sur un ensemble de règles de conversion de graphèmes en phonèmes [8] [4]. Ces règles ne prennent pas en compte les confusions phonémiques fréquentes chez les personnes dys et ne peuvent donc être utilisées telles quelles dans le cadre de la correction orthographique pour les personnes dys.

Sitbon propose de combiner l'utilisation du correcteur orthographique Aspell à celle du phonétiseur LIA_PHON [4] pour réécrire des phrases saisies par des enfants dys dans un moteur de recherche [18]. Dans un premier temps, la phrase est corrigée par Aspell pour obtenir des suggestions de correction, puis ces suggestions sont phonétisées avec LIA_PHON.

Nous avons adopté une approche différente qui intègre, dès la phase de transcription graphèmes-phonèmes, les confusions phonémiques fréquentes chez les dys. Nous augmentons ainsi le nombre de mots candidats en conservant les mots dont la forme phonétique n'est qu'approximativement celle du mot saisi.

3 Description du comparateur phonétique dissymétrique

Nous proposons une méthode pour corriger des textes rédigés par des personnes dys en français. Cette méthode opère sur des mots isolés, sans prise en compte du contexte. Un module compare chaque mot écrit aux mots d'un dictionnaire en les ayant préalablement transformés en phonétique et en prenant en compte, côté texte saisi, les particularités des fautes des dys. Cette méthode permet d'obtenir pour un mot erroné une liste de corrections potentielles. Un schéma du dispositif est présenté à la figure 1.

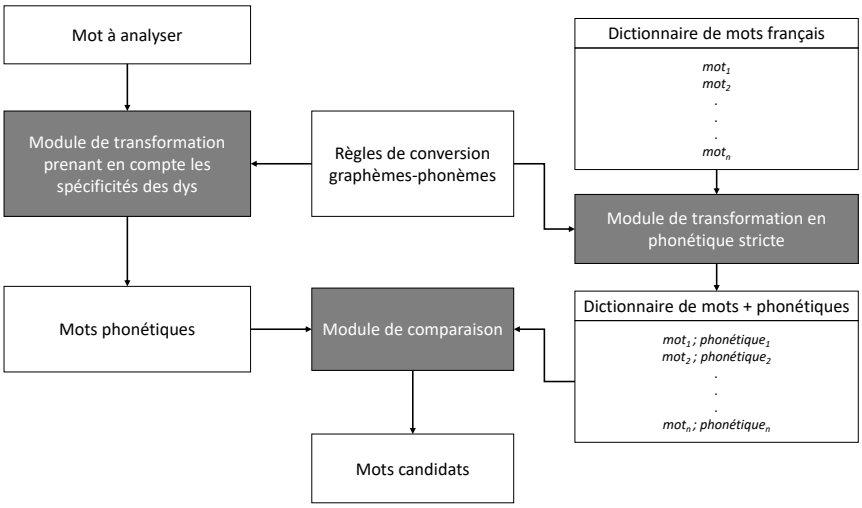

FIGURE 1 – Schéma du dispositif

3.1 Modules de transformation en phonétique

La première étape consiste à transformer les mots du dictionnaire et les mots du texte saisi en phonétique. Cette transformation est obtenue à l'aide d'algorithmes de transcription graphème-phonème exploitant un ensemble de règles. Ces algorithmes sont issus d'une application de

synthèse vocale pour aveugles. Les règles ont été simplifiées ou modifiées pour s'ajuster à notre problématique. Notamment, les marques de prosodie (rythme, accent tonique, etc.), qui ne sont pas utiles dans le contexte de la correction orthographique, ont été supprimées.

Règles de conversion graphèmes-phonèmes

Chaque règle permet de convertir un graphème en phonème en fonction du contexte (caractères précédents et suivants). Les règles ont le format suivant :

$CAR\_PREC\ [\ CAR\_ANALYSES\ ]\ CAR\_SUIV = PHON$
— CAR_PREC : caractères de contexte précédents
— $CAR\_ANALYSES$: caractères analysés (graphème)
— CAR_SUIV : caractères de contexte suivants
— $PHON$: transcription phonétique des caractères analysés (phonème)

Par exemple, la règle $f[ai]s\# = /ø/$ signifie que le graphème *ai* précédé de la lettre *f* et suivi de la lettre *s* et d'une voyelle correspond au phonème /ø/ (*eu* dans *peu*).

Pour encoder les phonèmes, nous utilisons une version de l'alphabet phonétique ARPABET [10] que nous avons adaptée pour le français. ARPABET utilise une représentation des phonèmes sur un ou deux caractères. Dans la version que nous utilisons, les consonnes sont codés avec un seul caractère, les voyelles avec deux caractères. Pour les voyelles, cela permet de représenter leur proximité phonétique sur le trapèze vocalique. Ainsi, les voyelles /i/ (dans *pris*) et /e/ (dans *pré*), qui diffèrent uniquement par leur degré d'ouverture (/i/ est plus fermé que /e/), correspondent respectivement à *IY* et *IH*. De plus, nous ne conservons que 11 voyelles sur les 13 à 16 recensés dans le système vocalique français selon les auteurs [5] [6] : par exemple, les voyelles /a/ et /ɑ/ (respectivement dans *patte* et *pâte*), dont l'opposition tend à disparaitre, ne constituent qu'une seule et même voyelle.

La plupart des règles n'ont qu'une transcription phonétique possible. Si deux règles entrent en contradiction, elles sont toutes les deux exploitées, ce qui créée un processus d'ouverture aux mots candidats.

Pour chaque graphème, il existe potentiellement plusieurs règles qui sont hiérarchisées des plus spécialisées au plus générales. Pour les mots saisis uniquement, viennent s'ajouter les règles prenant en compte les fréquentes confusions entre phonèmes proches (par ex. : les paires de consonnes /b/ et /p/, /v/ et /f/).

Module de transformation des mots du dictionnaire

La phase de transformation des mots du dictionnaire en phonétique respectent les règles phonétiques propres au français. Le module balaie les règles ordonnées jusqu'à la première applicable. Chaque mot du dictionnaire possèdent une, au maximum deux transcriptions phonétiques possibles (pour les homographes hétérophones comme *couvent* « *Les poules du couvent couvent.* »).

Module de transformation des mots saisis

Pour les mots saisis par la personne dyslexique, le module de transformation prend en compte l'ensemble des règles qui s'appliquent et retourne donc une liste de mots phonétiques possibles, en ne conservant que ceux qui sont plausibles. Par exemple, pour *fesé* les phonétiques possibles sont :

— fUHzIH correspondant aux mots *faisait*, *faisais*, *faisaient*
— fEHsIH correspondant aux mots *fessée*, *fessées*, *fessait*, etc.
— vEHsIH correspondant aux mots *vessait*, *vessais*, *vessaient*, etc.

3.2 Module de comparaison

Nous appliquons la distance de Damerau-Levenshtein entre les mots phonétiques résultants de la transformation phonétique des mots saisis et ceux du dictionnaire. La distance maximum choisie est de 1. Nous retenons non seulement tous les mots homonymiques mais également les mots dont la différence est de seulement 1 formant (par exemple /o/ dans *tôt* et /u/ dans *tout*).

4 Conclusion et travaux futurs

Nous avons présenté une méthode de comparaison fondée sur la phonétique et la modélisation des erreurs les plus fréquentes pour obtenir une liste de candidats potentiels à des mots erronés écrits par des dys. L'association d'une transcription phonétique prenant en compte les confusions grapho-phonémiques fréquentes des dys et d'une mesure de similarité permet d'obtenir de meilleures suggestions de corrections pour certaines des fautes des personnes dys. Cet algorithme a été testé sur quelques corpus et permet dès à présent de récupérer des mots que les correcteurs classiques ne trouvent pas (par exemple : « fesé » (faisais), « ésituron » (hésiterons), « osi » (aussi)).

Un travail approfondi d'amélioration est en cours grâce à la mise en place d'une distinction mots courts/mots longs et par le développement, également en cours, d'un module d'analyse contextuelle, celui-ci tenant compte d'un environnement potentiellement erroné du mot analysé. Ces travaux seront évalués sur l'ensemble des corpus à notre disposition et les résultats comparés avec les méthodes classiques de correction orthographique.

Références

[1] K. Atkinson. GNU Aspell. http ://aspell.net/.

[2] V. Bacquelé. 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), Oct. 2015.

[3] P. Barrouillet, C. Billard, M. D. Agostini, J.-F. Démonet, M. Fayol, J.-E. Gombert, M. Habib, M.-T. L. Normand, F. Ramus, L. Sprenger-Charolles, and S. Valdois. Dyslexie, dysorthographie, dyscalculie : bilan des données scientifiques. page 863, 2007.

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

[5] F. Carton. *Introduction à la phonétique du français.* Dunod edition, 1974.

[6] N. Catach. *L'orthographe française.* Nathan edition, 1986.

[7] F. J. Damerau. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3) :171–176, Jan. 1964.

[8] M. Divay and A. J. Vitale. Algorithms for Grapheme-phoneme Translation for English and French : Applications for Database Searches and Speech Synthesis. *Comput. Linguist.*, 23(4) :495–523, Dec. 1997.

[9] M. Habib. Bases neurologiques des troubles spécifiques d'apprentissage. *Réadaptation*, (486) :16–28, 2002.

[10] A. Klautau. ARPABET and the TIMIT alphabet, 2001.

[11] G. Kuenning. International Ispell. https ://www.cs.hmc.edu/ geoff/ispell.html.

[12] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8) :707–710, Feb. 1966.

[13] A. Mazur-Palandre. La dyslexie à l'âge adulte : la persistance des difficultés orthographiques. *SHS Web of Conferences*, 46 :10003, 2018.

[14] J. Pedler. Computer spellcheckers and dyslexics—a performance survey. *British Journal of Educational Technology*, 32(1) :23–37, 2001.

[15] L. Philips. Hanging on the Metaphone. *Computer Language*, 7(12 (December)), 1990.

[16] L. Philips. The Double Metaphone Search Algorithm. *C/C++ Users J.*, 18(6) :38–43, June 2000.

[17] R. Russel. INDEX (Soundex Patent), 1918.

[18] L. Sitbon, P. Bellot, and P. Blache. Eléments pour adapter les systèmes de recherche d'information aux dyslexiques. *Traitement Automatique des Langues*, 48(2) :123–147, 2008.