Conception et développement d'un module de traitement des lemmes comportant des lettres dérivatives

Cynthia Rakotoarisoa

Conception et développement d'un module de traitement des lemmes comportant des lettres dérivatives

Cynthia
RAKOTOARISOA

Sous la direction de Catherine BRISSAUD et Claude PONTON

Laboratoire : LIDILEM

UFR LLASIC
Département Sciences du Langage

Mémoire de master 2 mention Sciences du Langage - 20 crédits

Parcours : Industrie de la langue

Année universitaire 2020-2021

Conception et développement d'un module de traitement des lemmes comportant des lettres dérivatives

Cynthia
RAKOTOARISOA

Sous la direction de Catherine BRISSAUD et Claude POTON

Laboratoire : LIDILEM

UFR LLASIC
Département Sciences du Langage

Mémoire de master 2 mention Sciences du Langage - 20 crédits

Parcours : Industrie de la langue, orientation professionnelle

Année universitaire 2020-2021

Remerciements

Je souhaiterais tout d'abord remercier Claude Ponton pour sa confiance et son aide durant ce stage, mais également pour ces 2 années de Master.

Je remercie ensuite tous mes camarades de classe en particulier Ning, sans qui mes années de Master n'auraient pas été aussi agréables, Sanda, avec qui je passe toujours de bons moments et qui est là quand on a besoin, et Louis, pour le soutien moral et son humour.

Finalement, je remercie ma famille : mon père, et tous les trajets nocturnes qu'on a dû effectuer durant ces 6 dernières années, ma mère, pour sa présence même de loin, et mon frère, qui me fait toujours sortir de ma bulle de travail.

DÉCLARATION ANTI-PLAGIAT

1. Ce travail est le fruit d'un travail personnel et constitue un document original.
2. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
3. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
4. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).
5. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

PRENOM : CYNTHIA

NOM : RAKOTOARISOA

DATE : 02/09/2021

Sommaire

Introduction

Ce rapport présente mon travail réalisé dans le cadre de mon stage de Master 2 au sein du projet E-CALM qui s'intéresse aux performances des élèves à l'écrit en français tout au long du cursus scolaire. Mon travail porte sur les compétences en orthographe et plus particulièrement sur la question des lettres dérivatives muettes. Il s'agit de pouvoir observer et décrire les réussites et les difficultés des élèves face à ces lettres muettes en fin de mot. Pour cela, l'objectif principal de ce stage est, dans un premier temps, de concevoir un outil de détection de lettres dérivatives et muettes. Ensuite, à partir des résultats de cet outil, il sera possible d'observer les différents comportements des élèves face aux lettres finales muettes qu'elles soient dérivatives ou non.

La première partie de ce mémoire sera consacrée au cadre du stage et à la présentation du projet E-CALM. La seconde partie sera consacrée au travail réalisé durant ce stage et présentera le procédé ainsi que tout le processus de réflexion pour la conception de l'outil. Ensuite, on observera les compétences des élèves au niveau de l'orthographe des lettres muettes ainsi que leur évolution au fur et à mesure du cursus scolaire.

Partie 1

-

Cadre du stage

Chapitre 1. Projet E-CALM

1. Présentation

Mon stage s'effectue dans le cadre du projet ANR Écriture scolaire et universitaire : Corpus, Analyses Linguistiques, Modélisations didactiques (E-CALM[1]). Le projet consiste à décrire les compétences des élèves et étudiants à l'écrit (orthographe et cohérence textuelle) selon leur âge et les caractéristiques de leur milieu et également à comprendre les attentes des enseignants en fonction de leurs interventions sur les copies afin qu'elles puissent devenir des outils d'aide à l'écriture. Ce projet pluridisciplinaire réunit des spécialistes de sociologie du langage, de linguistique de corpus, d'analyse textuelle outillée et de didactique de l'écriture. Il s'organise au sein de quatre laboratoires :

- Le CIRCEFT[2] (référente : Elise Vinel) avec l'équipe ESCOL, constituée de linguistes et de sociologues du langage. Leurs travaux visent à étudier les inégalités sociales dans le cadre scolaire dont le rôle du langage dans l'acquisition du mode de fonctionnement et des exigences du système éducatif selon les contextes sociaux.

- Clesthia[3] (référente : Claire Doquet), expert dans le domaine de la constitution de corpus, développe le corpus ECRISCOL (cf. c. Corpus). Les chercheurs sont spécialisés dans l'étude de la génétique textuelle dans les écrits scolaires, l'acquisition de l'orthographe et l'évaluation des écrits des élèves. Clesthia a participé au développement de nombreux logiciels, notamment Lexico3 et Le Trameur, permettant une approche quantitative des textes.

- Le CLLE[4] (référente : Lydia-Mai Ho-Dac) développe une approche quantitative de la linguistique, à travers la constitution et l'exploitation de grands corpus langagiers, écrits ou oraux comme le corpus Resolco (cf. c. Corpus). L'équipe ERSS (Équipe de Recherche en Syntaxe et Sémantique) regroupe des spécialistes de l'analyse du niveau discursif, de l'acquisition et de l'enseignement du français et du traitement automatique des langues.

- Le LIDILEM[5] (référente : Marie-Paule Jacques) développe les corpus Littératie avancée et Scoledit (cf. c. Corpus). Les chercheurs effectuent des travaux sur la constitution de corpus, la description et la modélisation linguistique, la didactique des langues à l'écrit et le TAL.

1 http://e-calm.huma-num.fr/le-projet/
2 https://circeft.fr/escol/
3 http://www.univ-paris3.fr/clesthia-langage-systemes-discours-ea-7345-98241.kjsp
4 https://clle.univ-tlse2.fr/
5 https://lidilem.univ-grenoble-alpes.fr/

2. Objectifs

Les enquêtes PISA (OCDE 2016) ont démontré l'impact du contexte socio-économique sur les performances des élèves à l'école et également leur faible niveau dans des tâches d'écritures complexes. Cependant, dû au manque de données exploitables, il est difficile de déterminer plus précisément les performances des élèves. Le but du projet E:CALM est de décrire ces performances durant toute la scolarité jusqu'à l'université ainsi que les attentes des enseignants en termes de normes linguistiques. Le projet consiste notamment en la publication d'un vaste corpus d'écrits scolaires disponible en libre accès accompagné de métadonnées et d'outils de traitement.

Le projet a trois objectifs :

- Objectif 1 : structurer et mettre à disposition de la communauté scientifique un vaste corpus d'écrits d'élèves et d'étudiants;

- Objectif 2 : caractériser ces écrits et les attentes des enseignants du point de vue de l'acquisition de l'orthographe et de la cohérence, dans des analyses sociologiquement contextualisées;

- Objectif 3 : étudier les modalités d'écriture dans les avant-textes (plans, notes, brouillons) et les textes, notamment à travers l'influence réciproque des écrits remis et des interventions des enseignants sur les copies.

Pour réaliser ces objectifs, le projet est divisé en 7 tâches :

- Tâche 1 - Structuration d'un corpus cohérent et significatif (responsable : C. Ponton)

- Tâche 2 - Analyse des écrits scolaires et académiques : orthographe grammaticale et lexicale (responsable : C. Brissaud)

- Tâche 3 - Analyse des écrits scolaires et académiques : cohérence discursive (responsable : J. Rebeyrolle)

- Tâche 4 – Analyse des interventions écrites des enseignants sur les écrits des élèves (responsable : E. Bautier)

- Tâche 5 - Résultats et interprétation (responsable : C. Delarue-Breton)

- Tâche 6 – Exploitations du corpus à des fins de formation et d'enseignement (responsable : C. Garcia-Debanc)

- Tâche 7 - Diffusion et valorisation des ressources produites (responsable : S. Fleury)

3. Corpus

Le corpus E-CALM est composé de productions écrites d'élèves et d'étudiants. Il rassemble 4 corpus dont certains préexistants au projet. Ces ressources se différencient par le mode de recueil (exercice en classe ou bien exercice spécifique donné par les chercheurs), les consignes, les niveaux scolaires et s'il y a des interventions de l'enseignant sur les copies ou non.

EcriScol[6] (Doquet *et al*., 2017) regroupe 1500 textes écrits de l'école primaire à l'université afin d'étudier les écrits mais également les avant-textes (notes, brouillons, etc.). Ces données sont écologiques à savoir qu'il s'agit de travaux menés en classe sans intervention des chercheurs. La spécificité du corpus EcriScol est la prise en compte des traces d'écriture puisque la plupart des copies est associée à son brouillon et autres écrits préalables (notes, plans).

Littéracie Avancée[7] (Jacques et Rinck, 2017) contient 338 textes produits par des étudiants de niveau Licence 1 à Master 2. Les textes recueillis sont des écrits universitaires et professionnels (mémoire, lettre de motivation, etc.). Le but de ce corpus est de permettre d'analyser les compétences écrites et de déterminer les difficultés de l'écrit à un niveau avancé. Littéracie Avancée est le seul des quatre corpus contenant des textes dactylographiés.

Le projet Scoledit[8] (Wolfarth et al., 2017) a permis de collecter un corpus longitudinal de 3365 écrits du CP au CM2. Il a été demandé aux mêmes élèves de produire un texte narratif à partir d'images à différents moments de leur scolarité, ce qui permet de mettre l'accent sur l'évolution personnelle des compétences à l'écrit de chaque élève. L'analyse de toutes ces productions est effectuée grâce à des méthodes de TAL permettant d'aider à l'exploitation du corpus et d'observer les différents phénomènes présents dans les écrits.

ResolCo[9] (Garcia-Debanc *et al*., 2017) est un corpus de 400 manuscrits d'élèves et d'étudiants. La consigne donnée aux élèves était de rédiger un texte en utilisant trois phrases prédéfinies. Le but étant d'étudier l'organisation du discours et la cohésion textuelle selon les différents niveaux d'éducation.

6 http://www.univ-paris3.fr/ecriscol-300509.kjsp
7 https://lidilem.univ-grenoble-alpes.fr/ressources/corpus/litteracie-avancee
8 http://scoledit.org/scoledit/
9 http://redac.univ-tlse2.fr/corpus/resolco.html

Une fois collectées, toutes ces données passent par plusieurs étapes de traitement afin de pouvoir les exploiter numériquement.

Tout d'abord, les copies sont scannées et anonymisées. Ensuite les textes sont transcrits manuellement et encodés au format XML selon un guide de transcription développé dans le projet E:Calm. La transcription est effectuée de manière à respecter au maximum la production originale de l'élève, c'est-à-dire en gardant les erreurs, les ponctuations, les traces de révisions, etc. Les textes sont ensuite annotés manuellement pour proposer une version normalisée orthographiquement afin de pouvoir appliquer des outils TAL. Dans le processus de normalisation, les erreurs sont corrigées (orthographe, segmentation, etc.), mais tous les choix des temps verbaux faits par l'élève sont conservés. La ponctuation est également corrigée dans certains cas seulement : lorsqu'une ponctuation finale est suivie d'un mot commençant par une minuscule et inversement ou lorsqu'une énumération ne comporte pas de virgule.

Enfin, les données sont passées dans l'outil AliScol (Wolfarth, 2019) qui effectue un étiquetage morphosyntaxique avec TreeTagger[10] et une transcription phonétique associée à chaque forme avec l'outil LIA-PHON[11]. L'outil segmente également les productions en tokens pour aligner chaque forme produite avec leur forme normée, ce qui permet mener des analyses comparatives entre les deux.

10 https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
11 http://pageperso.lif.univ-mrs.fr/~frederic.bechet/download.html

Chapitre 2. Problématique

Le travail de ce stage se situe dans la tâche 2 du projet E-CALM. Le but de cette tâche est de décrire les compétences à l'écrit des élèves en analysant leurs réussites et leurs difficultés en orthographe. Trois aspects sont particulièrement ciblés : l'accord et la morphologie des verbes, l'accord et la morphologie des adjectifs, et la morphologie lexicale (lettres dérivatives).

Mon travail porte sur la question des lettres dérivatives pour les noms et adjectifs. Cela concerne par exemple, le -d de "grand" qui est utilisé dans "grandir" ou le -t de "chat" dans "chaton". En français, l'orthographe est un système qui favorise le phénomène des lettres muettes en fin de mot (Catach, 1995 ; Jaffré, 2005). Par conséquent, l'orthographe des lettres muettes pose particulièrement problème chez les apprenants du français (Fayol, Totereau, & Barrouillet, 2006; Sénéchal, 2000). Plusieurs questions se posent alors : quelles sont les difficultés rencontrées par les élèves concernant les consonnes muettes ? Est-ce que le fait que ces consonnes finales soient dérivatives a un impact sur leur orthographe ?

L'objectif de ce stage est de pouvoir assister automatiquement la description de l'orthographe des mots contenant ces lettres muettes et/ou dérivatives sur un grand ensemble de données. Il s'agit ici d'utiliser des méthodes de TAL pour tenter d'établir une modélisation de ces lettres dérivatives, de les annoter automatiquement dans le corpus E-Calm et d'analyser les erreurs/réussites des élèves. Ceci nous permettra également de voir dans quelle mesure le TAL peut assister les linguistes dans la description de ce phénomène et également les didacticiens à situer les performances des élèves et leur évolution.

Partie 2

-

Travail réalisé

Chapitre 4. Etude de l'existant

Durant ce stage, l'objectif est de concevoir un outil permettant de détecter automatiquement les mots contenant des lettres dérivatives afin d'assister le linguiste dans la description des erreurs sur cette catégorie de mots. Je vais commencer par explorer la littérature sur la question des lettres dérivatives muettes avant de voir quels problèmes elles posent aux élèves du point de vue orthographique. Je terminerai par présenter quelques ressources qui pourront être utiles pour concevoir notre outil.

1. Définition

Les lettres dérivatives ou morphogrammes contribuent en partie à la fréquence des lettres muettes en français. Les lettres muettes peuvent jouer un rôle grammatical et participer au marquage de la morphologie flexionnelle (le 's' muet du pluriel par exemple). Elles peuvent également faire partie de la morphologie dérivationnelle, c'est-à-dire qu'elles indiquent l'appartenance sémantique à une même famille de mot. Ces lettres seront donc prononcées dans les formes dérivées (par exemple, le 't' muet de "enfant" est prononcé dans "enfantin"). On retrouve également d'autres mots avec une lettre muette finale qui ne résulte pas de la morphologie dérivationnelle, mais qui possèdent une orthographe particulière (par exemple, le 'd' muet de "foulard" n'entraîne pas de formes dérivées).

Les lettres muettes sont des lettres présentent à l'écrit dans un mot, mais qui ne se prononcent pas à l'oral (par exemple, le 'p' de "trop" ou le 'c' de "blanc"). D'après Gingras et Sénéchal (2016), le problème de la définition des lettres muettes se pose pour deux types de fin de mots. D'abord, les unités finissant par un digramme ou un trigramme (ex. "ballet"). Pour certains, la consonne finale peut être considérée comme muette (Catach, 1995). Puis, les unités finissant par un 'e' précédé d'une consonne (ex. "place"), le 'e' peut être considéré comme une lettre muette (Peereman *et al*., 2018). Dans ces cas là, "-et" et "-ce" dans "ballet" et "place" sont considérés comme des graphèmes complexes[12]

Les lettres dérivatives sont des consonnes (muettes ou non) en fin de mot que l'on retrouve dans les formes dérivées de ce mot. Ces lettres dérivatives peuvent entraîner des formes dérivées provenant de la même famille morphologique (par exemple, chant-chanter). Ces lettres peuvent également être flexionnelles, c'est-à-dire qu'elles servent à produire des

12 Graphème composé de deux ou trois lettres, contrairement au graphème simple composé d'une seule lettre

formes fléchies, en général le féminin (par exemple, charmant-charmante). Le consonne finale d'un mot être dérivative, flexionnelle ou les deux à la fois.

2. Difficultés de l'orthographe des lettres muettes

D'après Sénéchal (2018), l'apprentissage de l'orthographe peut poser davantage de difficultés si «les correspondances phonèmes-graphèmes sont inconsistantes, instables, et conséquemment, peu prévisibles». C'est le cas du français qui est une langue morphophonologique (Ziegler *et al*., 2010). C'est pourquoi la majorité des mots en français se terminent par une lettre muette. Dans la base de données Silex (Gingras et Sénéchal, 2016), on trouve 56% de mots comportant une consonne finale muette dans les corpus d'enfants et 61% dans les corpus d'adultes (Gringas et Sénéchal, 2017). Les apprenants du français y sont donc fréquemment exposés, et ce depuis l'enfance.

Pour orthographier un mot, les jeunes apprenants s'appuient notamment sur la phonologie, en assemblant les phonèmes d'un mot aux graphèmes correspondant (Mousty et Leybaert, 1999), les lettres muettes peuvent donc être plus difficiles à apprendre. Jubenville, Sénéchal et Malette (2014) ont montré dans une étude avec des enfants de 8 ans que l'omission de la lettre finale muette représentait 95% de leurs erreurs d'orthographe. A 10 ans, l'orthographe de la terminaison muette semble toujours poser des difficultés par rapport aux mots sans consonne muette avec une fréquence d'apparition semblable dans les manuels scolaires (Sénéchal, 2018).

La production des lettres muettes en fin de mot peut être facilitée grâce à un apprentissage implicite (Treiman et Kessler, 2014). A travers la lecture, par exemple, les enfants peuvent rencontrer des régularités orthographiques et les appliquer à l'écrit (Danjou et Pacton, 2009). Dans une étude (Sénéchal, Gingras et L'Heureux, 2016), des enfants de 6 à 8 ans devaient épeler soit des mots terminant par la lettre muette 't', assez fréquente, soit par la lettre muette 'd', moins fréquente. Le 't' muet était plus souvent correctement orthographié que le 'd' muet.

Certaines consonnes muettes sont dues à la morphologie dérivationnelle, la sensibilité des élèves à la dérivation peut alors faciliter l'orthographe de celles-ci. Il a été montré dans deux études (Sénéchal, 2000; Sénéchal, Basque et Leclaire, 2006) que les mots avec une consonne finale dérivative seraient plus faciles à orthographier que les mots terminant par une lettre non dérivative. Les enfants de 9 ans avaient, en effet, de meilleurs résultats lorsque la lettre finale relevait de la morphologie (le 't' de "petit") plutôt que lorsqu'il s'agissait d'une orthographe spécifique au mot (le 's' de "fois"). Plus de 80% des erreurs sur la lettre muette

étaient l'omission ou la substitution. Whissell et Sénéchal (2017) ont également trouvé que les élèves réussissent plus souvent à produire correctement la lettre muette lorsqu'elle est dérivative contrairement aux mots sans dérivés.

Les élèves peuvent également utiliser des astuces et s'aider des relations morphologiques pour faciliter l'orthographe. Pour 75% des mots, les enfants de 10 à 11 ans ont affirmé avoir utilisé une stratégie consistant à les épeler (Ruberto, Daigle, et Ammar, 2016). En se servant des relations morphologiques, les élèves peuvent en déduire la présence d'une lettre muette et également de quelle lettre il s'agit. En effet, 75% des enfants de 9 ans se servent au moins une fois d'un mot dérivé pour déterminer la consonne muette finale, cette stratégie morphologique a permis de produire une orthographe correcte pour 77% des mots (Sénéchal *et al*., 2006).

3. Ressources disponibles

ManulexMorpho

ManulexMorpho[13] (Peereman, Sprenger-Charolles et Messaoud-Galusi, 2013) est une base de données lexicale développée dans le but de pouvoir analyser les erreurs à l'écrit, notamment la consistance des associations graphème-phonème (GP) et phonème-graphème (PG), en prenant en compte l'information morphologique. La base de données contient près de 10 000 mots apparaissant dans les manuels scolaires français à l'école élémentaire.

Dans ManulexMorpho, les graphèmes et phonèmes sont marqués par des indices morphologiques qui se trouvent en majorité en fin de mot. En effet, ces marques morphologiques posent souvent des difficultés à l'écrit, car elles sont souvent muettes à l'oral mais marquées à l'écrit ("e" final pour le genre; "s", "x", "ent" pour les pluriels des noms ou verbes). La base de données fournit ces informations en distinguant quatre catégories d'indices morphologiques :

1) les flexions de genre et de nombre (le "s" de "photos" ou le "e" de "brutale"), marquées par un "3",

2) les flexions verbales ("ent" dans "accordent"), marquées par un "4",

3) le morphographème "ent" des adverbes se terminant par "ment" est marqué par un "7",

4) les marques flexionnelles ou dérivationnelles de fin de mot ("s" de "gris"), annotées par un "6".

13 https://lpnc.univ-grenoble-alpes.fr/resources/ronald_peereman/Manulex_morpho/indexfr.html

Dans le Tableau 1, les lettres muettes sont représentées par un '#', elles sont notées avec un "6" lorsqu'elles sont dérivatives.

Orthographe	Phonologie	Catégorie grammaticale	Segmentation graphémique	Segmentation phonologique	Association G-P
blanc	bl@	ADJ	.b.l.an.6c	.b.l.@.6#	(b-b.l-l.an-@.6c-6#)
blanc	bl@	NC	.b.l.an.6c	.b.l.@.6#	(b-b.l-l.an-@.6c-6#)
chat	Sa	NC	.ch.a.6t	.S.a.6#	(ch-S.a-a.6t-6#)
doigt	dwa	NC	.d.oi.g.6t	.d.wa.#.6#	(d-d.oi-wa.g-#.6t-6#)
fois	fwa	NC	.f.oi.s	.f.wa.#	(f-f.oi-wa.s-#)
foulard	fulaR	NC	.f.ou.l.a.r.d	.f.u.l.a.R.#	(f-f.ou-u.l-l.a-a.r-R.d-#)
temps	t@	NC	.t.em.6p.s	.t.@.6#.#	(t-t.em-@.6p-6#.s-#)

Tableau 1: Extrait de la base ManulexMorpho

Silex

Silex[14] (Gingras et Sénéchal, 2016) est une base de données créée afin de faciliter l'évaluation des performances à l'écrit, notamment en ce qui concerne les lettres muettes en fin de mot. Silex est issu de deux bases de données, Manulexinfra[15] (Peereman *et al*., 2007) et Lexique 3.803[16] (New et al., 2004), et contient au total 119 664 mots. La base de données comprend un ensemble de classeurs Excel téléchargeables :

- Stimuli Selector : permet de sélectionner des mots en fonction de diverses statistiques et caractéristiques

- Table Generator : contient les fréquences et les distributions de probabilités des lettres muettes en fin de mot, des graphèmes correspondant à des phonèmes spécifiques et des lettres en fin de mot

- Master File : contient des informations destinées aux experts souhaitant contribuer ou personnaliser la base de données

Pour le traitement des lettres muettes, les mots se terminant par un digramme ou trigramme ou par un 'e' sont donc considérés comme n'ayant pas de lettre muette finale, cette option peut être modifiée par les utilisateurs. Davantage d'explications sont disponibles dans le guide d'utilisation de Silex[17]. Enfin, les formes fléchies sont également exclues pour distinguer les lettres muettes spécifiques au mot (radis) et celles dues aux règles de grammaire (amis).

14 https://carleton.ca/cllr/silex/
15 http://www.manulex.org/fr/infra/request.html
16 http://www.lexique.org/
17 https://carleton.ca/cllr/silex/user-guide/

OrthForm	PhonForm	SyntClass	OrthEnding	FinLetRole
accord	akoR	NOM	d	SL
compas	k&pa	NOM	s	SL
droit	dRwa	ADJ-ADV-NOM	t	SL
époux	epu	NOM	x	SL
fleur	fl9R	NOM	#	PL

Tableau 2: Extrait de la base Silex

On retrouve dans la base Silex, la forme orthographique du mot avec leur transcription phonétique. Chaque homographe constitue une seule entrée, c'est pour cela que l'on peut retrouver différentes catégories grammaticales dans la colonne "SyntClass". La colonne "OrthEnding" indique quelle est la consonne finale muette. Si le mot ne contient pas de consonne finale muette, on retrouve l'information '#'. Enfin, il est également possible d'obtenir uniquement le rôle de la lettre finale ("SL" si c'est une lettre muette ou "PL" si elle est prononcée).

Polymots

Polymots[18] (Gala et Rey, 2008) est une ressource lexicale de 19 510 mots qui présente des mots du français en groupe en fonction de leur structure morphologique. Elle a été construite manuellement à partir du Petit Larousse 2000. La base lexicale permet de visualiser la structure d'un mot (sa base et ses affixes), son type ('transparent', 'opaque' ou 'alternant'), sa productivité (nombre de membres dans la famille) et ses unités de sens (traits sémantiques). Elle peut être utile pour différencier les mots comportant une marque dérivative (lit) et ceux n'en comportant pas (brebis). Cette ressource est construite en fonction de la morphologie, la phonologie et l'étymologie. Elle a également été affinée (Gala *et al*., 2011) afin de caractériser sémantiquement les familles morphologiques et regrouper les mots partageant la même racine et une continuité de sens.

Lorsque la forme de base est de type transparent (TSP), le radical de ces mots a un sens. C'est-à-dire qu'il peut être utilisé comme un mot. Par exemple, les mots "effiler", "filature" et "filaire" ont tous comme forme de base "fil" qui est un mot qui existe.

Si la forme de base est de type alternant (ALT), les formes construites à partir de cette base entraîne une variation phonique du radical sans en modifier le sens. Par exemple, le /f/ de la forme "actif" peut devenir /v/ dans ses formes dérivées ("activité", "active", "réactivité", etc).

18 https://polymots.huma-num.fr/

Pour les mots avec une forme de base opaque (OPAK), le radical n'a pas de valeur sémantique. La base "voc", par exemple, a plusieurs formes dérivées ("vocal", "évoquer", etc), mais n'a pas de signification en tant que mot.

Base (type)	**Mots dérivés**
divin (TSP)	divin, divinateur, divinatoire, divinement, divination, diviniser, divinité
lit (TSP)	alitement, aliter, délitage, délitement, déliter, délitescent, lit, litée, literie, litière
bed (OPAK)	bedaine, bedon, bedonnant, bedonner
calc (OPAK)	calcaire, calcification, calcifié, calcique, calcium, calculabilité, calculable, calculateur, calculatrice, ...

Tableau 3: Exemples de familles morphologiques dans Polymots

Chapitre 5. Conception et réalisation du module

L'objectif du module est, pour chaque forme, d'ajouter les annotations suivantes : la consonne finale ('r', 't', 's', 'd', etc), si cette lettre est muette ou non (0 ou 1) et s'il s'agit d'une consonne dérivative (ou flexionnelle) ou non (0 ou 1). Lors de nos premières observations, une question s'est rapidement posée sur le fait de faire la distinction ou non entre les marques flexionnelles et les marques dérivatives. Il semble cependant difficile de séparer les deux, car la désinence du féminin (*éléphant-e*) et les affixes (*éléphant-eau*) ne semblent pas être fondamentalement différents d'un point de vue paradigmatique (Lehmann et Martin-Berthet, 2000). La consonne finale sera alors traitée de la même façon, qu'elle soit une marque dérivative ou flexionnelle.

1. Définitions

Avant de passer à la conception du module, il nous a fallu dans un premier temps définir ce que nous considérons comme lettre dérivative et comme lettre muette pour pouvoir établir des règles de modélisation. Ces définitions concerneront uniquement les noms et adjectifs dont le lemme se termine par une consonne.

Une lettre est muette lorsqu'elle est présente à l'écrit mais non prononcée à l'oral. Un mot contient donc une consonne finale muette lorsque la dernière lettre du lemme ne participe pas à la réalisation du dernier phonème du mot. A noter qu'une consonne finale faisant partie d'un digramme ou d'un trigramme n'est pas considérée comme muette (ex. ball**et**, resp**ect**, sol**eil**, n**ez**).

En ce qui concerne la lettre dérivative, je me suis appuyée sur la définition utilisée dans les travaux de ManulexMorpho. D'après Peereman, Sprenger-Charolles et Messaoud-Galusi (2013), une lettre dérivative est une consonne finale, qui peut être muette ou non. Cette consonne finale doit être prononcée dans les formes fléchies et/ou dérivées du mot (ex. le 'd' de *grand* pour donner *grande* ou *grandeur* ou le 's' de *anglais* pour donner *anglaise*). Cela prend en compte également les cas où il y a un changement d'orthographe par rapport à cette consonne ('c' *vs* 'ch' dans *blanc-blanche* ou 'x' *vs* 's' dans *époux-épouse*) et les cas de dénasalisation de la voyelle (*brun-brune*, *marin-marine*).

En explorant le corpus, certains cas n'entraient pas dans la définition de ManulexMorpho. A partir de cette définition, j'y ai alors ajouté quelques précisions :

- la consonne dérivative peut également être un élément de digramme ou de trigramme (*soin-soigner*, *secret-secrète*);

- il faut aussi y ajouter les cas de changement de lettre : une consonne est dérivative même si les formes dérivées entraînent un changement de consonne, que ce soit pour les lettres dérivatives "indirectes" (Catach, 1995), par exemple *amoureux-amoureuse*, ou bien pour les lettres "contradictoires" (Catach, 1995) du type *dissous-dissoute* qui sont plus marginales.

Il est à noter que les formes dérivées ou flexionnelles peuvent appartenir à la même catégorie syntaxique ou non.

Une fois avoir bien défini ce qu'est une lettre dérivative, je vais pouvoir appliquer cette définition pour la conception du module. Dans cette partie je présenterai les différentes étapes et les interrogations rencontrées durant le développement de l'outil.

2. Données

La détection des lettres dérivatives se fera donc sur les noms et adjectifs de la base de données E-CALM. J'ai extrait pour cela l'ensemble des noms et adjectifs dont le lemme se termine par une consonne pour les niveaux scolaires suivants : du CP au CM2, la 6ème et la 3ème. Il y a au total 48101 noms (dont 31428 dont le lemme se termine par une consonne) et 9293 adjectifs (dont 5902 avec un lemme comportant une consonne finale). En observant le Tableau 4, on peut avoir une idée du nombre de noms et d'adjectifs par niveau et de ceux qui se terminent par une consonne.

Niveau	NOM	% NOM_{cons}	ADJ	% ADJ_{cons}
3EME	4827	52%	1084	63%
6EME	5693	54%	1029	59%
CE1	5184	73%	933	69%
CE2	8856	71%	1585	64%
CM1	10274	68%	1986	60%
CM2	11242	65%	2217	62%
CP	2025	75%	459	86%

Tableau 4: Répartition des noms et adjectifs dans la base E-CALM

Tout d'abord il a fallu diviser ces données en 3 sous corpus : un corpus de test, un corpus de travail et un corpus d'évaluation. Pour obtenir ces 3 corpus, j'ai extrait pour chaque niveau environ 10% des formes parmi les noms et adjectifs. Dans le cadre de notre approche empirique, le corpus de test (environ 3400 formes) servira tout d'abord à observer sur un petit échantillon les différents cas de lettres muettes dérivatives, à avoir une estimation de leur nombre et de pouvoir également établir une ébauche pour le développement de l'outil. Le corpus de travail représente environ 80% du corpus des noms et adjectifs dont le lemme contient une consonne finale, il servira à affiner le modèle et à développer le module final. Finalement, le module sera appliqué sur le corpus d'évaluation (3400 formes) pour pouvoir mesurer sa performance en termes de précision et de rappel.

3. Phase d'observation

Cette phase d'observation consiste à jouer manuellement le rôle du module d'annotation sur le corpus de test afin de dégager les différents cas possibles et d'imaginer leurs traitements. J'ai aussi pu tester et affiner ma définition sur les lettres dérivatives muettes.

Dans le fichier sur lequel j'ai travaillé, on retrouve, pour chaque forme produite de chaque élève, sa forme normée avec son lemme, sa catégorie grammaticale, son genre ('m' pour masculin et 'f' pour féminin) et son nombre ('s' pour singulier et 'p' pour pluriel). On a également le type d'erreur observé sur la forme produite, qui a été calculé par AliScol (Wolfarth, 2019).

Niv	SegNorm	SegTrans	Categorie	Lemme	StatutErreur	Genre	Nombre
CM2	absent	absent	ADJ	absent	01-Normé	m	s
3EME	an	an	NOM	an	01-Normé	m	s
CE1	blanc	blan	ADJ	blanc	02-Phono	m	s
CM1	cachot	cacho	NOM	cachot	02-Phono	m	s
6EME	enfants	enfant	NOM	enfant	02-Phono	m	p
CE2	gens	gens	NOM	gens	01-Normé	_	p

Tableau 5: Extrait du corpus de test

Je me suis également interrogée sur certains cas, notamment lorsque la consonne finale change dans les formes dérivées. En effet, d'après la définition que l'on peut trouver dans ManulexMorpho (cf. 1. Définitions), même si la consonne finale est modifiée dans les formes dérivées, elle est tout de même dérivative, mais uniquement lorsqu'elle partage une même valeur phonétique (pour le concept de valeur phonétique voir Blanche-Benveniste et Chervel, 1969) que la lettre par laquelle elle a été remplacée. Par exemple, dans "choix", le 'x' est bien une consonne dérivative, car dans ses formes dérivées elle est remplacée par 's' (choisir) qui partage avec 'x' la valeur phonétique /z/. Ce cas de changement de lettre se distingue des cas où la consonne finale est remplacée par une lettre "contradictoire" (loup-louve, jus-juteux), qui n'est pas toujours accessible aux élèves. Pour le moment, nous avons pris la décision que tous les cas de changement de lettre seront traités comme consonne dérivative.

Dans le corpus, on trouve également certains mots avec deux consonnes muettes finales. On observe tout d'abord les mots où les deux consonnes muettes sont prononcées dans les formes dérivées (instinct-instinctif). Ceux-ci ne posent pas de problème dans l'annotation des données car les deux consonnes agissent de la même manière. Cependant, on trouve aussi certains mots qui finissent par deux consonnes muettes agissant différemment dans les formes dérivées. D'un côté, nous avons le cas où la consonne finale est bien prononcée dans les formes dérivées, mais pas l'avant dernière consonne qui peut rester muette (*vingt-vingtaine*). De l'autre, la consonne finale n'est pas dérivative, mais celle d'avant oui (*corps-corpulence*). Même si ce sont des cas intéressants, j'ai décidé de les traiter de la même manière que le reste du corpus, c'est-à-dire en me focalisant uniquement sur la consonne finale. Pour les mots "*vingt*" et "*instinct*", la consonne finale 't' est donc muette et dérivative, pour "*corps*" la consonne finale 's' est muette et non dérivative. Lorsqu'un mot contient deux consonnes finales muettes, l'information n'est pas présente dans le corpus annoté, mais ce mot est bien repéré par le programme final.

Un autre problème rencontré est dû au fait que la prononciation de certaines lettres finales varie en fonction des locuteurs. Par exemple, certains mots acceptent deux prononciations : sourcil, persil, ananas, fait, but, etc. Pour l'annotation j'ai décidé de choisir la prononciation considérée comme la plus fréquente, c'est-à-dire celle indiquée dans Silex. Je ne prendrai donc pas en compte le fait que certaines consonnes finales peuvent être prononcées ou non.

Pour résumer, si la consonne finale du lemme est modifiée dans les formes dérivées, cette consonne sera tout de même considérée comme dérivative. L'annotation des consonnes dérivatives concernera uniquement la lettre finale du lemme, ce sera également le cas pour l'annotation des consonnes muettes, peu importe la nature de la lettre précédente. Enfin, pour le cas des mots acceptant plusieurs prononciations, la plus fréquente sera prise en compte.

Niv	SegNorm	SegTrans	Categorie	Lemme	StatutErreur	ConsFinale	ConsMuet	ConsDeriv
6EME	choix	choit	NOM	choix	02-Phono	x	1	1
CM1	corps	corps	NOM	corps	01-Normé	s	1	0
CE1	vingt	blan	ADJ	blanc	02-Phono	c	1	1
CE2	loup	loup	NOM	loup	01-Normé	p	1	1
6EME	ananas	ananas	NOM	ananas	01-Normé	s	0	0
CE2	gens	gens	NOM	gens	01-Normé	s	1	0

Tableau 6: Extrait du corpus de test annoté

Maintenant que le procédé d'annotation pour ces lettres dérivatives et muettes a été défini, on peut obtenir un aperçu de l'occurrence des lettres dérivatives muettes à partir du corpus de test annoté manuellement. Dans le Tableau 7, on peut voir, pour chaque type de consonne finale, lequel est le plus fréquent. On remarque qu'en grande majorité, la consonne finale sert à la dérivation et que dans 46% des cas elle est muette et dérivative.

	Muette	**Non muette**
Dérivative	1576 46%	1119 33%
Non dérivative	241 7%	455 13%

Tableau 7: Type de la consonne finale dans le corpus de test (3391 formes)

4. Annotation automatique

Une fois le corpus de test annoté manuellement, je vais pouvoir commencer à développer une première version du module pour l'annoter automatiquement. Les informations à ajouter seront donc : la consonne finale, si c'est une consonne muette et si c'est une consonne dérivative. En entrée, nous avons le corpus de test, un fichier au format .csv, les

ressources Silex et Manulex au format .csv et l'outil LIA_PHON. En sortie nous avons le corpus de test annoté avec les informations supplémentaires sur la consonne finale toujours au même format. Le module intégrant ce modèle sera ensuite entièrement développé avec le langage de programmation Python[19].

a. Consonnes muettes

Tout d'abord, j'ai commencé par travailler sur le traitement des consonnes muettes. En effet, je disposais de suffisamment de ressources pour les traiter sans trop de difficultés. Pour cela, j'ai utilisé la base de données Silex (cf. 1.b. Ressources), qui recense les unités du français à consonne finale muette, et l'outil LIA_PHON[20] (Béchet, 2001), qui est un système de phonétisation automatique qui fonctionne à base de règles. Dans la base Silex, on peut retrouver soit dans la colonne "OrthEnding" la ou les consonnes finales muettes ou bien dans la colonne "FinLetRole" le rôle de la lettre finale, c'est-à-dire "SL" si elle est muette ou "PL" si elle est prononcée (Tableau 8). LIA_PHON associe chaque graphème ou ensemble de graphèmes au phonème correspondant, représenté par deux caractères, en fonction du contexte. Pour les mots "instinct" et "jaloux", on obtiendra en sortie les alignements suivants : inin#sss#ttt#inin#ct# et jjj#aaa#lll#ouou#x#. Dans la sortie de LIA_PHON, le # sépare les couples graphème/phonème et les phonèmes sont codés sur les deux dernières lettres. Ainsi « jjj » désigne le graphème « j » associé au phonème « jj » (cf. Annexe 1 : Correspondances LIA_PHON/SAMPA).

OrthForm	OrthEnding	FinLetRole
abord	d	SL
doigt	gt	SL
public	#	PL
réveil	#	PL

Tableau 8: Extrait d'annotation des lettres finales muettes ou non dans Silex

L'utilisation de Silex a été relativement simple car l'information nécessaire sur la ou les lettres finales muettes est déjà disponible, notamment pour repérer les doubles consonnes muettes. Cependant, étant donné que la base Silex est issu de Manulex constitué à partir de manuels scolaires, tous les lemmes de la base E-CALM ne sont pas présents. J'avais donc

19 https://www.python.org/
20 http://pageperso.lif.univ-mrs.fr/~frederic.bechet/download.html

besoin d'une ressource complémentaire. Avec LIA_PHON, j'ai pu simplement créer un fichier texte contenant la phonétisation de tous les lemmes du corpus E-CALM. J'ai ensuite vérifié automatiquement le phonème associé au graphème final de chaque lemme pour déterminer si la consonne finale est muette ou non.

Cette partie du module donne *a priori* de très bons résultats concernant l'annotation automatique des lettres muettes puisque les résultats obtenus automatiquement correspondent bien à l'annotation manuelle effectuée précédemment. Ceci sera bien entendu à confirmer sur un corpus plus large.

b. Consonnes dérivatives

Le traitement des consonnes dérivatives a été un peu plus complexe que pour les consonnes muettes. Pour cette partie, j'ai utilisé la base ManulexMorpho (cf. 1.b. Ressources), où l'on retrouve l'information sur la dérivation de la consonne finale.

Comme indiqué dans le Tableau 4, la plupart des consonnes finales sont des marques de dérivation. Pour le développement du modèle, je suis donc partie du principe que, par défaut, toutes les consonnes finales étaient dérivatives et qu'il me fallait repérer celles qui ne l'étaient pas. Avant d'utiliser une quelconque ressource, j'ai commencé par essayer de repérer les consonnes dérivatives avec des expressions régulières. En effet, on peut retrouver des motifs réguliers au niveau de la fin de certains mots qui marquent une dérivation ou non. Par exemple, pour beaucoup de noms qui se finissent en -ment et qui sont dérivés de verbes, la lettre finale n'est pas dérivative (*déménager-déménagement*, *ranger-rangement*, etc). Il y a aussi les noms en -at comme "*commissariat*" ou "*orphelinat*" où le 't' final n'est pas dérivatif.

Cependant, en établissant uniquement des expressions régulières, on obtient un filtrage qui se base uniquement sur la forme du mot et sans traiter les autres caractéristiques linguistiques du mot. Par exemple, avec seulement les expressions régulières on ne peut pas différencier le « t » de "*chocolat*", qui est dérivatif, du « t » de "*commissariat*" ou de "*orphelinat*" qui le sont pas. Avec Silex, il est également possible d'avoir accès à d'autres caractéristiques linguistiques. Pour le cas de l'exemple précédent, j'ai utilisé Silex afin de déterminer si un nom peut également être utilisé comme adjectif, ce qui est le cas de "chocolat". Lorsqu'un nom peut également être utilisé comme adjectif, sa consonne finale sera automatiquement notée comme dérivative. En effet, les consonnes finales des adjectifs sont en grande partie dérivatives, notamment en raison des formes fléchies du féminin.

A partir de là, la plupart des cas les plus fréquents ont été traités. Ainsi, je suis passée au traitement de cas plus particuliers comme certains noms invariables. Dans Silex, l'information sur l'invariabilité d'un mot n'est pas disponible, mais il est possible de prédire cette information en observant les catégories grammaticales. Comme mentionné précédemment, il est possible avec Silex d'obtenir les différentes catégories grammaticales auxquelles appartient un mot. En effet, chaque entrée correspond à tous les homographes d'une orthographe. J'ai alors sélectionné les noms qui pouvaient également être utilisés comme préposition (ex. "*envers*") ou pronom (ex. "*rien*") et j'ai noté la consonne finale comme non dérivative.

Silex m'a également permis de repérer directement les mots comportant deux consonnes finales muettes. On peut remarquer que dans la majorité des cas des doubles consonnes silencieuses, la lettre finale n'est pas dérivative (*temps*, *puits*, *poids*) sauf lorsque le mot se termine par un 't' (*vingt*, *doigt*).

Pour compléter mon programme, j'y ai ajouté la base ManulexMorpho, qui contient environ 64% des lemmes du corpus des noms et adjectifs dont le lemme se termine par une consonne. Les lemmes les plus fréquents seront donc bien traités, cependant plusieurs consonnes n'étaient pas notées comme dérivatives alors qu'elles correspondaient bien à la définition établie. J'ai donc pris en compte cette éventualité où un lemme serait considéré comme sans consonne dérivative par ManulexMorpho alors que c'est le cas. Dès lors que la lettre finale d'un mot sera considérée comme non dérivative par la base ManulexMorpho, ce mot sera passé dans le programme décrit précédemment afin de vérifier si c'est bien le cas ou non.

Orthographe	Phonologie	Catégorie gram-maticale	Segmentation graphémique	Segmentation phonologique	Association G-P
bonbon	b§b§	NC	.b.on.b.on	.b.§.b.§	(b-b.on-§.b-b.on-§)
charmant	SaRm@	ADJ	.ch.a.r.m.4ant	.S.a.R.m.4@	(ch-S.a-a.r-R.m-m.4ant-4@)
direction	diREksj§	NC	.d.i.r.e.c.t.i.on	.d.i.R.E.k.s.j.§	(d-d.i-i.r-R.e-E.c-k.t-s.i-j.on-§)
ouvert	uvER	ADJ	.ou.v.e.r.4t	.u.v.E.R.4#	(ou-u.v-v.e-E.r-R.4t-4#)
sort	soR	NC	.s.o.r.t	.s.o.R.#	(s-s.o-o.r-R.t-#)
robot	Robo	NC	.r.o.b.o.t	.R.o.b.o.#	(r-R.o-o.b-b.o-o.t-#)

Tableau 9: Extrait des entrées ne possédant pas de consonne dérivative dans ManulexMorpho

En comparant les résultats obtenus automatiquement avec ceux annotés manuellement, j'ai calculé le rappel et la précision de mon modèle. Sur l'ensemble du corpus de test avec 485 lemmes uniques, on obtient une précision[21] de 0.84 et un rappel[22] de 0.96. Les cas restants sont principalement dus à certaines expressions régulières qui peuvent être assez générales et prendre en compte beaucoup de mots. Par exemple, avec les mots se terminant en -eur, on retrouve principalement des noms comme "*menteur*", "*serveur*" ou "*chauffeur*" qui ne produisent pas de formes dérivées ou flexionnelles avec le 'r' final. Cependant, parmi ces noms finissant en -eur on peut retrouver des cas avec une consonne dérivative, comme "*fleur*", "*chaleur*" ou "*peur*". Ce résultat appuie l'idée que ce phénomène de lettre dérivative peut être traité automatiquement et que l'on peut retrouver certaines similarités entre les mots possédant une consonne finale dérivative.

5. Modèle final

Une fois avoir développé ce premier modèle, je l'ai appliqué sur le corpus de travail qui contient 29774 noms et adjectifs dont le lemme contient une consonne finale (1481 lemmes uniques). A partir du corpus de travail annoté, j'ai vérifié quelques annotations. J'ai ensuite pu affiner ce premier modèle avec les nouveaux cas qui n'étaient pas présents dans le corpus de test. Je n'ai pas eu à faire de modifications majeures par rapport au premier modèle, j'ai principalement ajouté ou modifié quelques expressions régulières, par exemple pour les mots se finissant par -ieur (ex. "*ingénieur*", "*supérieur*") et pour les mots se finissant par -ier,

21 La précision est la proportion d'éléments correctement classés par rapport à l'ensemble des éléments classés.
22 Le rappel est la proportion d'éléments correctement classés par rapport à l'ensemble des éléments de cette classe.

même si on retrouve beaucoup de cas où le 'r' final est dérivatif (exemple : "*sorcier*", "*infirmier*"), cela ne prend pas en compte d'autres cas fréquents où le 'r' n'est pas dérivatif (ex. "*escalier*", "*collier*"). On obtient alors les données présentées dans le Tableau 10 avec le modèle final.

	Muette	**Non muette**	**Total**
Dérivative	14748 50%	9442 32%	24190
Non dériva-tive	2068 7%	3516 12%	5584
Total	16816		29774

Tableau 10: Types de consonne finale dans le corpus de travail

Dans le tableau suivant, j'ai également regroupé les consonnes finales muettes les plus fréquentes dans ce corpus. On note une proéminence du 't' en position finale, suivi du 's' qui, dans beaucoup de cas, n'est pas dérivatif (*fois*, *dessus*, *concours*, etc.)

Consonne finale	**Dérivative**	**Non dérivative**
d	739	3
p	2190	0
s	664	1682
t	10238	151
x	445	153

Tableau 11: Répartition des consonnes finales muettes fréquentes dans le corpus de travail

Voici un récapitulatif de l'algorithme général conçu pour repérer les lettres dérivatives et muettes dans le corpus E-CALM.

Pour chaque lemme du corpus :

- on récupère la consonne finale,

- on détermine si la dernière lettre est muette ou non (avec Silex et LIA_PHON),

- on utilise ManulexMorpho pour annoter les lettres dérivatives,

- on applique les règles du Tableau 12 pour les lemmes qui n'ont pas été annotés par ManulexMorpho afin de savoir si la consonne finale est dérivative ou non.

Consonne dérivative	Consonne non dérivative
• NOM qui peut également être un ADJ → *gagnant*, *quotidien*, *câlin* • expressions régulières → mots se terminant par : -ent, -oix, etc → *parent*, *choix* • dans la plupart des cas	• mots composés → *arc-en-ciel*, *chauve-souris* • double consonne muette sauf lorsque la consonne finale est 't' → *corps* • mots "invariables" → *rien*, *envers* • expressions régulières → fin de mot en : -euil, -ement, -at, etc → *fauteuil*, *déguisement*, *orphelinat*

Tableau 12: Règles générales du modèle de détection des lettres dérivatives

Evidemment, ces règles peuvent ne pas fonctionner pour certains cas particuliers, pour cela j'ai établi une liste d'exceptions dans mon algorithme lorsque la règle ne s'appliquait pas. Cette liste peut toujours être complétée et modifiée. Le module n'a pas pu être évalué avec le corpus d'évaluation, car la constitution de ce "gold corpus" était trop coûteuse en temps.

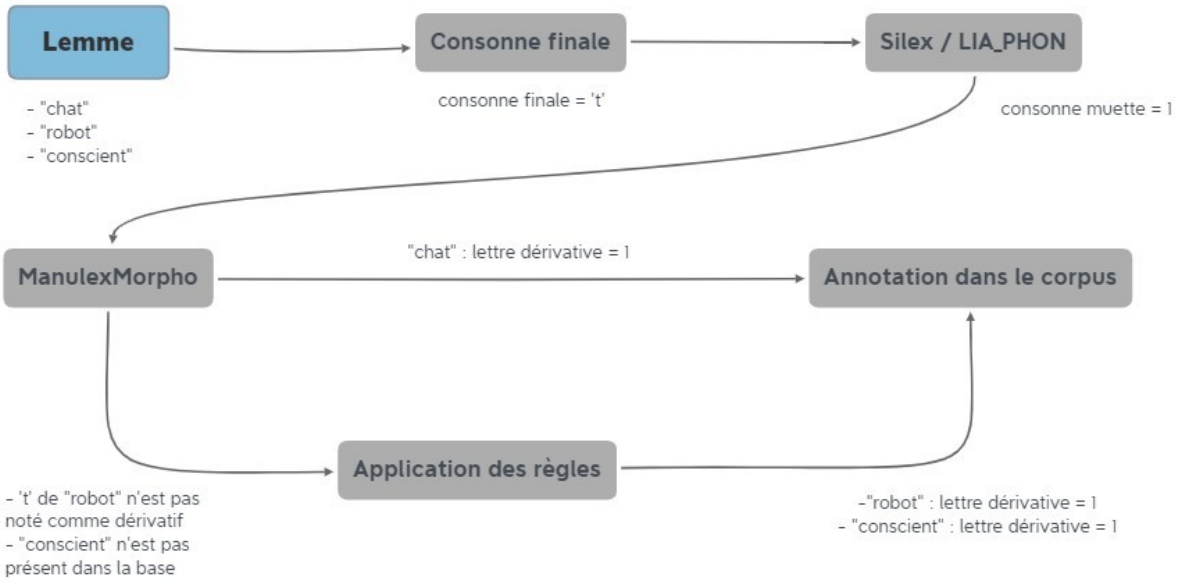

Figure 1: Chaîne de traitement de détection des lettres dérivatives et muettes

Chapitre 7. Analyse des productions

Une fois l'ensemble des données du corpus traité, il faut analyser les productions des élèves et étudier les erreurs commises sur ces mots.

J'ai tout d'abord observé les erreurs dans leur globalité, c'est-à-dire que j'ai sélectionné toutes les formes non normées, peu importe le type d'erreur qui a été calculé par AliScol. On remarque dans le Tableau 13, que lorsque la consonne finale est muette, les élèves ont tendance à faire moins d'erreurs sur les mots où la consonne finale est dérivative.

	Muette	**Non muette**
Dérivative	19,23% (2830)	20,94% (1449)
Non dérivative	22,49% (465)	22,65% (792)

Tableau 13: Répartition des erreurs en fonction du type de consonne finale

Cependant, même si le taux d'erreur est moins important pour les formes comportant une consonne finale muette et dérivative que pour les formes sans consonne dérivative, ce résultat est encore à approfondir. De plus, on peut voir que pour la catégorie des consonnes finales non dérivatives, le taux d'erreur est sensiblement le même que la consonne finale soit muette ou non. Il serait donc intéressant d'observer les erreurs au niveau de la consonne finale afin d'obtenir plus de précisions sur ces erreurs en particulier.

1. Découpage des noms

Pour pouvoir faire une comparaison plus détaillée entre les formes produites par les élèves et les formes normées, il a fallu développer un algorithme de découpage des noms en [radical+flexion de nombre] (par exemple, *élèves* => *élève* + *s*). Ce travail permettra de distinguer la consonne finale (dérivative ou non) à la marque du pluriel. Par exemple, si l'élève a écrit "*robots*" au lieu de "*robot*", il faut pouvoir déterminer que l'erreur ne se situe pas sur la consonne finale 't' qui est bien orthographiée, mais qu'il s'agit d'une erreur de nombre. Ce cas doit également être différencié du cas où l'élève produit "robos" au lieu de "robot", ici on retrouve bien une erreur sur la consonne finale.

Ce découpage sera basé sur le même principe que AliAdj (Gaubil, 2020), un module de traitement conçu pour le découpage des adjectifs dans le corpus E-CALM. Le principe d'AliAdj est de proposer un découpage des formes normées et des formes produites des adjectifs selon un modèle linguistique afin de déterminer où se trouvent les erreurs produites sur les adjectifs. Ce découpage se présente sous cette forme : base + flexion de genre + flexion de nombre. Voici dans le Tableau 14 quelques exemples de modèles et de découpages pour les adjectifs.

Modèles	Découpage			
	Masculin		Féminin	
	Singulier	Pluriel	Singulier	Pluriel
absent	absent + _ + _	absent + _ + s	absent + e + _	absent + e + s
agile	agile + _ + _	agile + _ + s	agile + _ + _	agile + _ + s
additionnel	additionnel + _ + _	additionnel + _ + s	additionnell + e + _	additionnell + e +s
affreux	affreux + _ + _		affreus + e + _	affreus + e + s
complet	complet + _ + _	complet + _ + s	complèt + e + _	complèt + e + s
copain	copain + _ + _	copain + _ + s	copin + e + _	copin + e + s
doux	doux + _ + _		douc + e + _	douc + e+ s
égal	égal + _ + _	éga + _ + ux	égale + e + _	égal + e + s
faux	faux + _ + _		fauss + e + _	fauss + e + s
favori	favori + _ + _	favori + _ + s	favorit + e + _	favorit + e + s
fou	fou + _ + _	fou + _ + x	foll + e + _	foll + e + s

Tableau 14: Extrait des découpages des adjectifs (Gaubil, 2020)

Puisque tous les découpages des adjectifs sont disponibles dans la base E-CALM, je vais pouvoir m'en servir directement pour l'analyse des erreurs sur les consonnes finales des adjectifs. Pour les noms, le découpage se fera tout simplement sous la forme : [base + flexion de nombre]. Dans la grande majorité des noms, la base correspond au lemme et la flexion du nombre à la marque du pluriel 's'. Pour les cas plus particuliers comme "animaux" ou "travaux", je les ai découpés de cette façon : anima + ux et trava + ux. Même si dans le contexte des lettres dérivatives, le découpage de ces cas n'a pas grande importance, car la consonne finale du lemme est remplacée par la marque du pluriel. Enfin, pour les noms dont

le lemme et la forme du pluriel est identique, on gardera le mot entier comme base afin que la consonne finale ne soit pas considérée comme marque du pluriel. J'ai alors développé un script rapide pour le découpage automatique des noms à partir de leur forme normée.

			Découpage	
Segnorm	**Genre**	**Nombre**	**Base**	**Nombre**
abricots	m	p	abricot	s
loup	m	s	loup	
animaux	m	p	anima	ux
gens	_	p	gens	
palais	m	_	palais	

Tableau 15: Exemples de découpage des noms

Ce découpage va me permettre de me focaliser sur l'orthographe de la base et d'observer directement la consonne finale. Il faut maintenant pouvoir comparer ces formes normées découpées aux formes produites.

2. Erreurs sur la consonne finale

Afin de décrire les erreurs produites sur les consonnes finales, nous avons décidé de les classer en plusieurs catégories :

- 01- Normé : la consonne finale est bien orthographiée (forme produite → 'cha**t**'),

- 02- Oubli : la consonne finale est omise (forme produite → 'cha_'),

- 03- Modifiée : la consonne finale n'est pas la bonne (forme produite → 'concour**t**'),

- 04- Ajout : la consonne finale est bien présente, mais l'élève a ajouté une lettre supplémentaire en position finale du mot (forme produite → 'bazar**d**').

Pour identifier ces types d'erreurs, il faudrait comparer automatiquement la base des formes normées à celle des formes produites pour retrouver la consonne finale. Pour les adjectifs, le découpage des formes produites est déjà disponible dans le corpus, il suffit donc

de vérifier que le dernier caractère de la base normée et de la base produite soit le même. Par exemple, si la forme normée est 'blancs' et que la forme produite est 'blanc', il y a une erreur de nombre, mais on retrouve bien la bonne consonne finale dans la forme produite.

SegNorm	SegTrans	BaseAdjNorm	BaseAdjTrans	StatutCons
amical	amicale	amical	amical	01- Normé
content	contamt	content	contamt	01- Normé
chaud	chau	chaud	chau	02- Oubli
costaud	costaut	costaud	costaut	03- Modifiée
court	coure	court	coure	03- Modifiée
long	longt	long	longt	04- Ajout

Tableau 16: Extrait des erreurs sur la consonne finale des adjectifs

Pour les noms, le procédé reste le même, cependant, étant donné que je ne dispose pas du découpage des formes produites des noms, j'ai effectué ce découpage en même temps que l'analyse de la consonne finale.

Tout d'abord, on retrouve les noms où la forme normée correspond au lemme. Dans ce cas-là, il suffit tout simplement de comparer les derniers caractères, de la même façon que pour les adjectifs, pour déterminer si la consonne finale a bien été orthographiée ou non et de quel type d'erreur il s'agit. Par exemple, avec la forme normée "*loup*", si la forme produite est "*lout*" la consonne finale est modifiée, par contre si la forme produite est "*lou*" il s'agit d'une erreur d'omission.

Ensuite, pour les noms au pluriel où la forme normée est différente du lemme, j'ai effectué l'analyse en partant de la droite. C'est-à-dire, que dans un premier temps, j'ai vérifié la marque du pluriel et dans un second temps, la consonne finale. Lorsque le dernier caractère de la forme produite n'est pas une marque du pluriel, on comptera une erreur de nombre (par exemple : "*biscuits*" écrit "*biscuit* "). On analysera ensuite la consonne finale de la même manière que décrite précédemment pour déterminer le type d'erreur. S'il n'y a pas d'erreur de nombre et que la marque du pluriel est bien présente, c'est ici qu'on effectuera un découpage base + nombre de la forme produite. Par exemple, avec la forme normée "*bords*", la forme produite "*bors*" sera découpée de cette façon : bor + s. On étudiera la base de la forme

produite afin de déterminer le type d'erreur sur la consonne finale. Dans l'exemple précédent, le type d'erreur dans "*bors*" est un oubli de la consonne finale 'd'.

Durant le traitement des erreurs, j'ai remarqué que dans certains cas où la consonne finale produite est un 's', le choix du type d'erreur sur la consonne finale peut poser problème. S'agit-il d'une marque du pluriel ou bien d'une consonne finale spécifique à l'orthographe du mot ? Cette question concerne particulièrement les cas où la forme normée est au singulier et ne contient pas de 's' final et que la forme produite contient un 's' final. Par exemple, lorsque la forme normée est "*chat*" et la forme produite est "*chas*", il y a plusieurs interprétations possibles, soit la forme produite est au pluriel avec un oubli de la consonne 't', soit la consonne finale a été modifiée et remplacée par 's'. J'ai alors décidé de procéder de la manière suivante : si la forme normée est au singulier, on considère que la forme produite est également au singulier peu importe la consonne finale produite. Dans l'exemple précédent, "*chas*" correspond à une modification de la consonne finale. Cependant, lorsque l'erreur correspond à l'ajout d'un 's' après la consonne finale, il s'agit d'un erreur de nombre. Par exemple, lorsque la forme normée est "*chat*" et que la forme produite est "*chats*", on considère que l'élève a fait une erreur de nombre mais a bien su orthographier la consonne finale.

Voici plusieurs exemples reprenant les différents cas cités précédemment :

SegNorm	SegTrans	SatutCons	ErrNombre
loup	lout	03- Modifiée	0
loup	lou	02- Oubli	0
buisson	buissons	01- Normé	1
travail	travaille	04- Ajout	0
biscuits	biscuit	01- Normé	1
bords	bors	02- Oubli	0
chat	chas	03- Modifiée	0
chat	chats	01- Normé	1
chats	chas	02- Oubli	0

Tableau 17: Extrait des erreurs sur les noms

3. Résultats

Une fois toutes les données traitées, il faut analyser les résultats obtenus et voir les difficultés que présentent les élèves face aux lettres muettes.

Observons tout d'abord ces résultats d'un point de vue global. Le graphique suivant montre la répartition des types d'erreur sur la consonne finale en fonction des différentes caractéristiques de la consonne finale, qu'elle soit muette (M), non muette (nM), dérivative (D) ou non dérivative (nD).

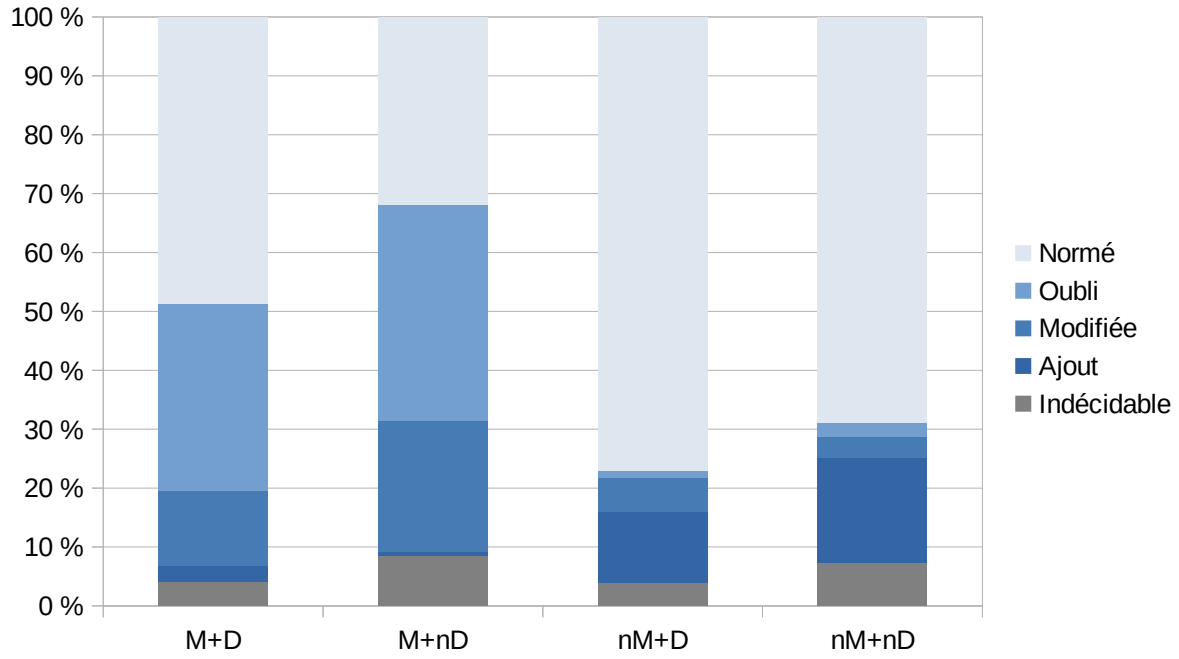

Figure 2: Répartition des types d'erreurs selon les caractéristiques de la consonne finale

Dans l'ensemble des données, on remarque une meilleure réussite de l'orthographe sur la consonne finale muette lorsque celle-ci est dérivative (environ 50% d'erreurs) que lorsqu'elle ne l'est pas (environ 68% d'erreurs). Le type d'erreur le plus fréquent est l'omission qui représente un peu plus de la moitié des erreurs commises sur la consonne finale, suivi des erreurs de modification de la consonne finale. On note également que les erreurs de modification de la consonne finale est plus importante lorsque le mot se termine par une lettre non dérivative.

Quant aux erreurs d'ajout de lettre après la consonne finale, elles sont plus fréquentes lorsque la consonne finale n'est pas muette. En effet, on retrouve de nombreuses fois l'ajout d'un 'e' final ("travail**le**", "animal**e**", etc) qui ne change pas la prononciation du mot. Les élèves peuvent également ajouter une consonne dérivative aux mots qui n'en contiennent pas ("bazar**d**", "cauchemar**d**", etc). Cela peut sous-entendre que les élèves s'appuient sur des formes dérivées des mots afin de déduire la présence d'une consonne finale ou non.

Les erreurs notées comme "Indécidable" sont des formes où le type d'erreur sur la consonne finale n'a pas pu être déterminée automatiquement. On retrouve des cas comme : "déguisemexx", "écroulmaie" ou "seuye" au lieu de "seuil" que je n'ai pas pu traiter, car on a une modification de la consonne finale 'l', mais également l'ajout d'une autre lettre en position finale. Ces erreurs indécidables représentent 5% des erreurs au total.

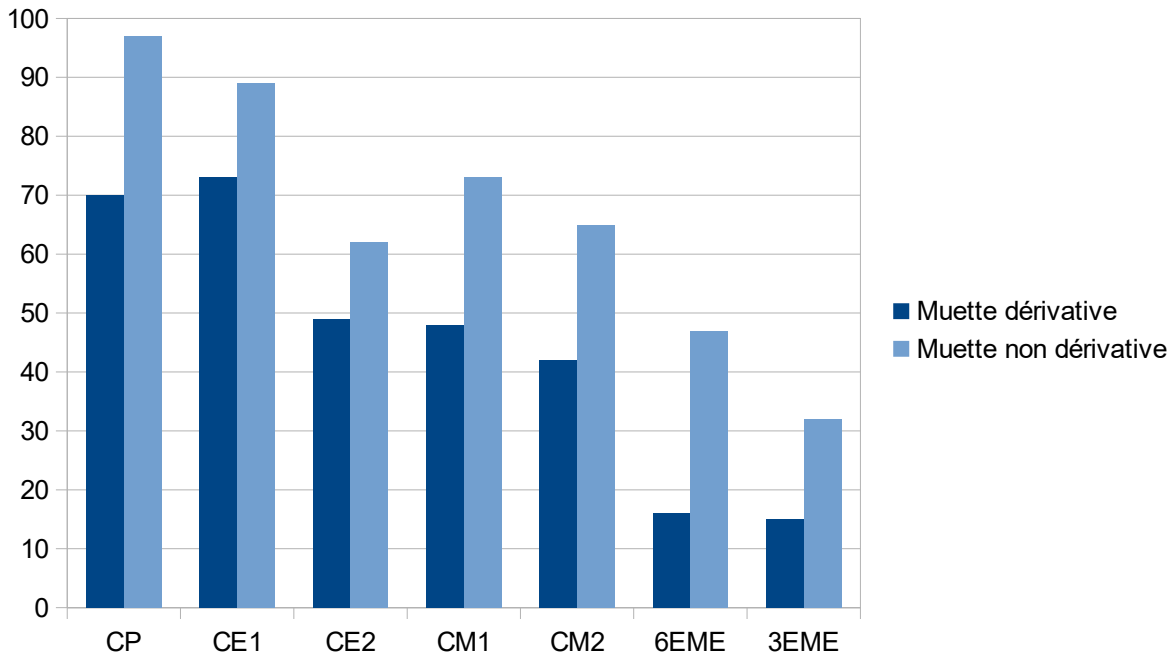

Figure 3: Taux d'erreurs sur les consonnes finales muettes par niveau

Ce schéma regroupe le taux d'erreurs sur les consonnes finales muettes selon les niveaux scolaires. On remarque tout d'abord une tendance générale à la baisse des erreurs sur la consonne finale. En effet, plus le niveau scolaire est avancé, moins les erreurs sur la consonne finale sont fréquentes, avec une moyenne d'environ 83% d'erreur sur la consonne finale muette en CP contre une moyenne de 25% pour les élèves de 3ème. L'orthographe des

consonnes dérivatives semble poser de moins en moins problème au fur et à mesure du cursus scolaire. On passe de 70% d'erreur sur la consonne finale dérivative en CP à 15% pour la 3ème. L'information sur la consonne finale dérivative devient donc plus accessible pour les élèves ayant un niveau scolaire plus élevé. Quant aux consonnes finales muettes non dérivatives, elles sont également mieux orthographiées dans les niveaux scolaire les plus avancés. Cependant, on remarque, que ce soit en 3ème ou en CE1, la consonne finale non dérivative est toujours plus difficile à produire que lorsqu'elle est dérivative.

Conclusion

1. Bilan

L'objectif de ce travail était de concevoir un outil de détection et d'annotation des lettres muettes et/ou dérivatives afin de pouvoir observer et décrire les difficultés des élèves sur les mots qui en comportent. La conception de l'outil est faite, même s'il reste encore beaucoup de choses à approfondir. En effet, j'ai traité le sujet uniquement dans sa globalité et il y a de nombreux phénomènes qu'il serait intéressant d'aborder. Je pense notamment à la distinction entre les différentes typologies de lettres dérivatives (changement de consonne, double consonnes muettes) que je n'ai pas eu le temps de traiter et le besoin d'une évaluation de l'outil qui permettrait son amélioration.

Les résultats obtenus montrent une différence des performances sur l'orthographe des lettres muettes en fonction de leur nature dérivationnelle ou non. On observe également une progression de l'orthographe des lettres muettes au fur et à mesure de la scolarité. Le système permettant de déterminer le type d'erreur sur la consonne finale peut encore être affiné, en effet il reste encore quelques cas qui ne sont pas encore traités par l'algorithme.

Le travail effectué sur les lettres dérivatives n'est évidemment pas terminé, ce stage a permis de faire un premier travail de mise en ordre sur la question et peut servir de base pour d'autres études plus approfondies sur les lettres muettes et dérivatives.

2. Bilan personnel

D'un point de vue personnel, ce stage m'a permis dans un premier temps de mettre en pratique ce que j'ai étudié tout au long de ce Master et d'intégrer mes compétences au sein d'un projet intéressant et original. J'ai également pu découvrir le monde de la recherche avec pas mal de documentation, mais également beaucoup de questions qui se posaient sans cesse au fur et à mesure du projet. J'ai également appris à adopter une méthodologie de travail différente de la mienne. Durant ce stage j'ai dû travailler en autonomie et organiser moi-même mon temps de travail en fonction des tâches données par mon tuteur chaque semaine. Même si tout au long du Master j'ai pu faire plusieurs travaux personnels, cela m'a poussée à avoir davantage confiance en moi-même concernant mon travail. En effet, étant habituée aux

travaux de groupe où les tâches étaient souvent réparties, je ressors satisfaite d'avoir pu concevoir un tel outil de A à Z et d'avoir pu également obtenir des résultats concrets.

Même si la situation sanitaire ne m'a pas permis de vivre l'expérience complète de ce stage, cela m'a permis de m'adapter selon différents lieux de travail (chez moi, à la bibliothèque, au campus) que j'ai pu varier selon mes envies et mes besoins.

Bibliographie

Doquet C. (2020) « Analyser linguistiquement l'écriture à l'école: EcriScol, un corpus génétique. » *Working Papers in Linguistics* Volume 4, 4, pp.127-140.

Jacques M.-P., Rinck F. (2017). « Un corpus de "littéracie avancée" : Résultat et point de départ ». *Corpus*, 16, pp.217-237. https://journals.openedition.org/corpus/2806

Wolfarth, C., Ponton, C., Totereau, C. (2017). « Apports du TAL à la constitution et à l'exploitation d'un corpus scolaire ». Dans Doquet C., David J. & Fleury S., Spécificités et contraintes des grands corpus de textes scolaires : problèmes de transcription, d'annotation et de traitement, *Corpus*, 16:2017, 185-214.

Garcia-Debanc, C., Ho-Dac L.-M., Bras, M. et Rebeyrolle, J.. (2017). « Vers l'annotation discursive de textes d'élèves ». *Corpus*, 16. Mis en ligne le 09 janvier 2018. URL : http://journals.openedition.org/corpus/2783

Wolfarth C., Ponton C., Totereau C.. (2017) «Apports du TAL à la constitution et à l'exploitation d'un corpus scolaire au travers du développement d'un outil d'annotation orthographique, Constitution d'un corpus scolaire et TAL.» *Spécificités et contraintes des grands corpus de textes scolaires : problèmes de transcription, d'annotation et de traitement*, 16, pp.185 - 214. ⟨hal-01878701⟩

Jaffré J-P. (2005) «L'orthographe du français, une exception ?»

Fayol, M., Totereau, C. & Barrouillet, P. (2006) «Disentangling the impact of semantic and formal factors in the acquisition of number inflections: Noun, adjective and verb agreement in written French.» https://doi.org/10.1007/s11145-005-1371-7

Hargrave, A. C., & Sénéchal, M. (2000). «A book reading intervention with preschool children who have limited vocabularies: The benefits of regular reading and dialogic reading.» *Early Childhood Research Quarterly*, 15, 75-90. http://dx.doi.org/10.1016/S0885-2006(99)00038-1

Sénéchal, M. (2018). «Comment les élèves apprennent-ils l'orthographe ?» *Cnesco. Écrire et rédiger : comment guider les élèves dans leurs apprentissages. Notes des experts.* https://www.cnesco.fr/fr/ecrire-et-rediger/

Gingras M., Sénéchal M. (2017) «Silex: A database for silent-letter endings in French words.»

Jubenville K., Sénéchal M., Malette M.. (2014) «The moderating effect of orthographic consistency on oral vocabulary learning in monolingual and bilingual children.» *Journal of Experimental Child Psychology*

Treiman R., Kessler B. (2014) «How Children Learn to Write Words.»

Ruberto, N., Daigle, D., Ammar, A. (2016). «The spelling strategies of francophone dyslexic students.» *Reading and Writing: An Interdisciplinary Journal* https://doi.org/10.1007/s11145-015-9620-x

Sénéchal M. (2006) «Testing the Home Literacy Model: Parent Involvement in Kindergarten Is Differentially Related to Grade 4 Reading Comprehension, Fluency, Spelling, and Reading for Pleasure.» *Scientific Studies of Reading*

Peeremanl R., Sprenger-Charolles L., Messaoud-Galusi S. (2013) «The contribution of morphology to the consistency of spelling-to-sound relations: A quantitative analysis based on French elementary school readers.»

New B., Pallier C., Brybaert M., Ferrand L. (2004) «Lexique 2 : A new French lexical database.»

Gala N., Rey V. (2008) «POLYMOTS : une base de données de constructions dérivationnelles en français à partir de radicaux phonologiques.»

Gala N., Hathout N., Nasr A., Rey V., Seppälä S. (2011) «Création de clusters sémantiques dans des familles morphologiques à partir du TLFi.»

Sitographie

E-CALM : http://e-calm.huma-num.fr/le-projet/
CIRCEFT : https://circeft.fr/escol/
Clesthia : http://www.univ-paris3.fr/clesthia-langage-systemes-discours-ea-7345-98241.kjsp
CLLE : https://clle.univ-tlse2.fr/
LIDILEM : https://lidilem.univ-grenoble-alpes.fr/
EcriScol : http://www.univ-paris3.fr/ecriscol-300509.kjsp
Littéracie Avancée : https://lidilem.univ-grenoble-alpes.fr/ressources/corpus/litteracie-avancee
Scoledit : http://scoledit.org/scoledit/
ResolCo : http://redac.univ-tlse2.fr/corpus/resolco.html
TreeTagger : https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
LIA_PHON : http://pageperso.lif.univ-mrs.fr/~frederic.bechet/download.html

Table des Tableaux

Table des Figures

Table des annexes

Annexe 1

Correspondances LIA_PHON/SAMPA

LIA_PHON	SAMPA	Exemples
ii	i	idiot, ami
ei	e	ému, été
ai	E	perdu, maison
aa	a	alarme, patte
oo	O	obstacle, corps
au	o	auditeur, beau
ou	u	coupable, loup
uu	y	punir, élu
EU	2	creuser, deux
oe	9	malheureux, peur
eu	@	petite, fortement
in	e~	peinture, matin
an	a~	vantardise, temps
on	o~	rondeur, bon
un	9~	lundi, brun
yy	j	piétiner, choyer
ww	w	quoi, fouine
pp	p	patte, repas, cap
tt	t	tête, net
kk	k	carte, écaille, bec
bb	b	bête, habile, robe
dd	d	dire, rondeur
gg	g	gauche, égal, bague
ff	f	feu, affiche, chef
ss	s	sœur, assez, passe
ch	S	chanter, machine, poche
vv	v	vent, inventer, rêve
zz	z	zéro, raisonner, rose
jj	Z	jardin, manger, piège
ll	l	long, élire, bal
rr	R	rond, chariot, sentir

mm	m	madame, aimer, pomme
nn	n	nous, punir, bonne
##	_	(silence marker)

Table des matières

MOTS-CLÉS : orthographe, lettres muettes, lettres dérivatives, TAL

RÉSUMÉ

L'une des difficultés les plus fréquentes lors de l'apprentissage du français est l'orthographe des lettres muettes qui sont très présentes à l'écrit. En effet, la majorité des mots en français se terminent par une lettre muette. Ces lettres muettes sont principalement dues à la morphologie dérivationnelle et marquent l'appartenance à une même famille de mots (le t de 'chat' dans 'chaton'). Ce mémoire présente le processus de la conception d'un outil de traitement automatique du langage permettant de détecter automatiquement ces lettres dérivatives. Cet outil permettra ensuite d'étudier les performances des élèves face aux lettres muettes selon leur nature dérivationnelle ou non.

KEYWORDS : spelling, silent letters, derivational letters, NLP

ABSTRACT

The spelling of silent letters is one of the most common difficulties for French learners. Moreover these silent letters are very common in French. Indeed, most of the words in French end with a silent letter. These silent letters are mainly due to derivational morphology and specify the belonging of the same semantic and morphological family (the t in 'chat' in 'chaton'). This master thesis introduces the process of designing a natural language processing tool to detect these derivational letters. This tool will then be used to study students' skills beside silent letters according to their derivational or non-derivational nature.