Outiller la description de la morphologie adjectivale du primaire à l'université

Rachel Gaubil

Outiller la description de la morphologie adjectivale du primaire à l'université

GAUBIL
Rachel

Sous la direction de Claude Ponton et Catherine Brissaud

Laboratoire : LIDILEM

UFR LLASIC
Département Sciences du Langage

Mémoire de master 2 Sciences du Langage – orientation Recherche – 20 crédits

Parcours : Industries de la Langue

Année universitaire 2019-2020

Outiller la description de la morphologie adjectivale du primaire à l'université

GAUBIL
Rachel

Sous la direction de Claude Ponton et Catherine Brissaud

Laboratoire : LIDILEM

UFR LLASIC
Département Sciences du Langage

Mémoire de master 2 Sciences du Langage – orientation Recherche – 20 crédits

Parcours : Industries de la Langue

Année universitaire 2019-2020

Remerciements

Je souhaiterais tout d'abord remercier chaleureusement M. Claude Ponton pour sa disponibilité et sa patience constantes tout au long de ce projet.

Je remercie également M. Olivier Kraif d'avoir accepté de faire partie de mon jury et de m'avoir appris à persévérer lors d'un précédent stage concernant le projet PhraseoRom.

Je tiens aussi à remercier toute l'équipe du projet E-Calm et principalement M. Jacques David et Mme Catherine Brissaud pour leur aide sur les aspects linguistiques ainsi que Mme Claire Wolfarth qui m'a fourni les données nécessaires à ce travail.

Un grand merci à ma famille, à Antoine pour ses encouragements, à Charlotte pour sa relecture et à mon équipe de choc du master : Justine, Lucie, Marie, Myriam, Vincent pour son soutien indéfectible.

Université Grenoble Alpes

DÉCLARATION

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.

NOM : GAUBIL

PRENOM : Rachel

DATE : 27/08/2020

Sommaire

Introduction

A l'ère où nos écrits sur téléphones, tablettes et ordinateurs se « corrigent » d'eux-mêmes grâce aux correcteurs automatiques, maitriser l'orthographe reste-t-il une nécessité ? Ces outils s'améliorent de jour en jour cependant ils ont toujours des défauts et restent, comme leur nom l'indique, des outils. Au-delà de l'imperfection de ces systèmes, il y a également tout un pan de notre quotidien qui demande des compétences en littératie. En effet, tout ne se fait pas au travers de ces correcteurs, prenons par exemple : les devoirs scolaires dans lesquels l'orthographe est presque toujours évaluée, les documents administratifs (formulaires, lettres, etc.) qui doivent être manuscrits, etc. Dans le monde du travail, rendre un document comportant des fautes d'orthographe peut être vu comme un manque de professionnalisme et porter préjudice. Par ailleurs, l'orthographe est souvent considérée comme un marqueur social et pour rester dans l'exemple du monde professionnel, c'est également un critère de sélection lors de candidatures pour un emploi. Ainsi, l'apprentissage de l'orthographe reste une des bases de la littératie dans notre société. Cette dernière est en effet définie par l'OCDE[1] comme « l'aptitude à comprendre et à utiliser l'information écrite dans la vie courante, à la maison, au travail et dans la collectivité en vue d'atteindre des buts personnels et d'étendre ses connaissances et ses capacités » (OCDE & Statistique Canada, 2000, p. 10).

Par ailleurs, depuis plusieurs années les enquêtes PISA[2] cherchent à évaluer les compétences des élèves et ont pu montrer qu'en France, le niveau était assez bas dans les tâches d'écritures et de compréhension de l'écrit complexes (OCDE, 2019). Cependant, les compétences réelles des élèves dans ce domaine ne sont que très peu décrites car nous manquons de données représentatives et ce à chaque niveau scolaire. C'est à cela que s'intéresse le projet E-Calm : Ecriture scolaire et universitaire : Corpus, Analyses Linguistiques, Modélisations didactiques. Il cherche en effet à décrire et analyser les productions des élèves du CP jusqu'au Master grâce au développement d'un vaste corpus de productions et d'outils numériques pour l'exploiter. Il vise également à fournir aux enseignants ainsi qu'à leurs formateurs des outils didactiques pour l'accompagnement et l'évaluation des élèves.

[1] Organisation de Coopération et de Développement Économiques
[2] Programme International pour le Suivi des Acquis des élèves

A cette fin, le projet propose de constituer et de mettre à disposition un large corpus d'écrits d'élèves et d'étudiants. C'est sur ce corpus que seront menées des descriptions de l'acquisition des normes orthographiques et de cohérence de texte à travers le filtre du contexte sociologique des écrits. Enfin, l'évolution des productions au cours de leur écriture sera également analysée afin de repérer quelle est l'influence des interventions des enseignants sur les textes des élèves et étudiants. Ces différentes phases se basent sur trois hypothèses[3] :

1. « La langue écrite des élèves et des étudiants peut être décrite comme intégrant les normes et les procédés scripturaux de différents niveaux : orthographe, morphosyntaxe, cohésion/cohérence et textualité. »
2. « Les ratures, en tant qu'indicateurs du travail des scripteurs sur leur texte en cours d'écriture, permettent de repérer des éléments propices à un enseignement efficient. »
3. « Les variations dans les performances sont liées aux variations des contextes sociaux, didactiques et pédagogiques (rapports à l'écriture, usages de l'écrit scolaire, variations interindividuelles). »

Mon mémoire s'inscrit dans le cadre de ce projet financé par l'ANR[4] et qui se déroule de 2018 à 2021. Le projet E-Calm est coordonné par Claire Doquet du laboratoire CLESTHIA à l'Université Paris 3.

Il regroupe quatre laboratoires dont un de Sciences de l'Education : CIRCEFT-ESCOL (Paris) et trois de Sciences du Langage : CLESTHIA (Paris), CLLE-ERSS (Toulouse) et LIDILEM (Grenoble). Chaque laboratoire de Sciences du Langage est responsable d'une partie des ressources de ce projet (cf. Tableau 1). En effet, il regroupe les corpus récoltés dans quatre autres projets : Scoledit, Resolco, Ecriscol et Littéracie avancée qui seront présentés plus explicitement dans la partie 1.1 Récolte des données (cf. Chapitre 1. Contexte et problématique, p. 12).

[3] http://e-calm.huma-num.fr/le-projet/ [consulté le 07/08/2020]
[4] Agence Nationale de la Recherche

Projet	**Laboratoire superviseur**	**Université**
Scoledit	LIDILEM	UGA
Resolco	CLLE-ERSS	Toulouse 2
Ecriscol	CLESTHIA	Paris 3
Littéracie avancée	LIDILEM	UGA

Tableau 1 : Laboratoires et Universités responsables de chaque projet

Les quatre laboratoires se partagent la responsabilité des sept tâches en lesquelles sont divisées le projet E-Calm ; le récapitulatif est disponible en Annexe 1 (p. 88).

Mon travail de recherche s'inscrit dans la tâche 2 de ce projet. Plus précisément, il s'intéresse à l'analyse des erreurs concernant l'adjectif, que ce soit au niveau grammatical ou lexical. Ce travail se place donc tout naturellement au sein du LIDILEM (Linguistique et Didactique des Langues Étrangères et Maternelles) qui est responsable de cette tâche. Le LIDILEM possède quatre axes principaux de recherche[5] :

1. Description et modélisation linguistiques, corpus, TAL[6] ;
2. Didactique des langues : analyse et évaluation des processus d'enseignement/apprentissage ;
3. Acquisition du langage : multimodalité, variabilité et contexte ;
4. Sociolinguistique : identités, cultures, interactions, usages.

Le projet E-Calm se situe à la croisée des axes 1 et 2. Par ailleurs, les membres du LIDILEM impliqués dans ce projet sont les suivants : Catherine Brissaud (PR[7]), Marie-Paule Jacques (MCF[8]), Claude Ponton (MCF), Fanny Rinck (MCF), Isabelle Rousset (IGR[9]), Corinne Totereau (MCF) et Claire Wolfarth (Docteure).

C'est principalement avec Claude Ponton et Catherine Brissaud que j'ai collaboré puisqu'ils sont les encadrants de mon stage. J'ai également travaillé avec Claire Wolfarth car c'est elle qui a conçu le système permettant le traitement des données recueillies et qui fournira les ressources d'entrée de mes recherches.

[5] https://lidilem.univ-grenoble-alpes.fr/node/16/axes-recherche [consulté le 24/08/2020]
[6] Traitement Automatique du Langage
[7] Professeur.
[8] Maître de Conférences.
[9] Ingénieur de Recherche.

Partie 1

-

Contexte du projet

Chapitre 1. Contexte et problématique

Comme expliqué précédemment, le travail présenté dans ce mémoire fait partie intégrante du projet E-Calm qui s'intéresse aux performances des élèves, que ce soit au niveau de l'orthographe ou de la cohérence discursive des écrits.

L'étude de ces performances et de leurs évolutions s'appuie sur un large corpus couvrant toute la scolarité jusqu'au Master et regroupant 6 754 textes, soit environ 4 937 700 mots (cf. Annexe 2, p. 90).

1. Ressources

La constitution de ces ressources s'est faite en s'appuyant sur les quatre projets cités précédemment : Scoledit, Resolco, Ecriscol et Littéracie avancée. C'est l'ensemble des données recueillies lors de ces travaux qui constitue le corpus E-Calm. Chacun de ces projets a été mené par une équipe différente et a suivi une méthodologie qui lui est propre. Ils étaient tous, mis à part Resolco, déjà en cours ou achevés avant le projet E-Calm. C'est pourquoi un protocole commun a été mis en place durant la construction du corpus *Resolco* et c'est lors du projet E-Calm que le travail de mise en cohérence des autres corpus a commencé. Je ne présenterai ici que Scoledit, Resolco et Ecriscol car dans le cadre de mon travail je n'ai pas eu accès au corpus Littératie avancée qui n'avait pas encore été finalisé. De même, mes données regroupent uniquement les productions d'élèves de CP, CE1, CE2, CM1, 3ème et 1ère année de licence car les autres niveaux étaient toujours en cours de traitement (transcription et normalisation) lorsque j'ai commencé mon travail.

1.1. Récolte des données

Le corpus Scoledit fait suite au projet Lire – Écrire au CP[10] (Goigoux, 2015) qui a récolté des productions d'élèves de CP et de CE1 pour mener une étude longitudinale en suivant les mêmes élèves d'une année à l'autre. Scoledit a été la continuité de cette étude jusqu'au CM2 même si pour cela le nombre d'élèves a dû être réduit. En effet, les membres de l'équipe du premier projet étaient plus nombreux et ils disposaient donc de beaucoup plus

[10] Cette recherche a été financée par l'Institut Français de l'Éducation, la DGESCO et le laboratoire ACTé de l'Université Clermont Auvergne.

de ressources humaines pour aller récolter les données dans les différentes classes. Au niveau du CP, il a été demandé aux élèves d'écrire l'histoire du petit chat à partir des quatre images ci-dessous (cf. Figure 1). Pour les niveaux du CE1 au CM2, les élèves devaient écrire un texte après avoir choisi un ou deux personnages parmi les quatre images qui leur ont été présentées (cf. Figure 2).

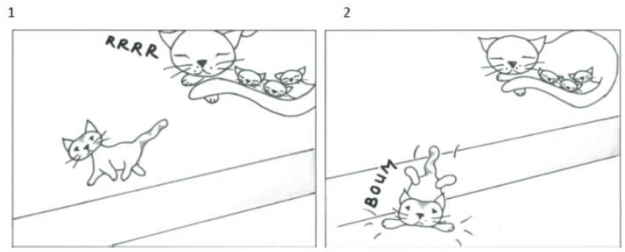

Figure 1 : Images présentées aux élèves de CP

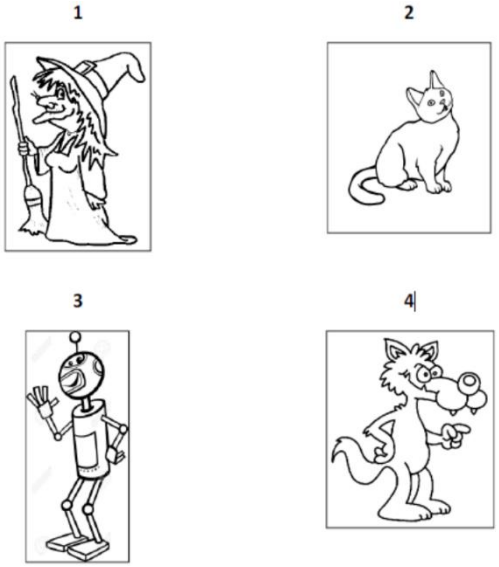

Figure 2 : Images présentées aux élèves du CE1 au CM2

Voici deux exemples de productions réalisées respectivement par un élève de CP[11] (cf. Figure 2) et un de CM1[12] (cf. Figure 4) en respectant les consignes décrites plus tôt. Elles ont été récupérées sur Scoledition[13], le site internet du projet Scoledit qui permet de visualiser l'ensemble du corpus de primaire.

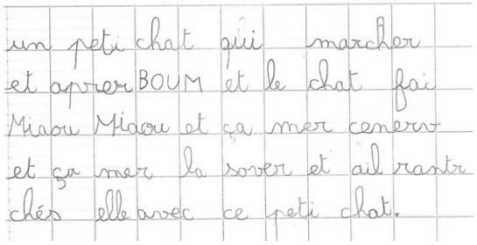
un peti chat qui marcher
et aprer BOUM et le chat fai
Miaou Miaou et ça mer coleres
et ça mer la soser et ail rastr
chés elle avec ce peti chat.

Figure 3 : Exemple de production d'un élève de CP

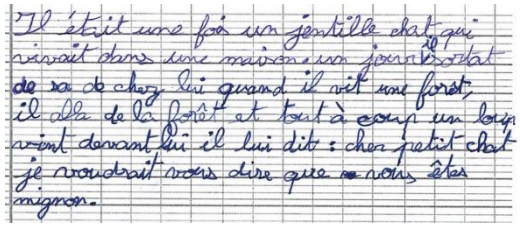
Il était une fois un jentille chat qui
vivait dans une maison. un jour il était
de sa de chez lui quand il vit une forêt.
il alla de la forêt et tout à coup un loup
vient devant lui il lui dit : chez petit chat
je voudrait vous dire que vous êtes
mignon.

Figure 4 : Exemple de production d'un élève de CM1

Le corpus Resolco, lui, concerne mes données de 3ème. La consigne vise à produire des écrits contenant des phénomènes de cohésion/cohérence. Pour cela, il a été demandé aux

[11] Corpus Scoledit, CP, élève 49
[12] Corpus Scoledit, CM1, élève 221
[13] Voir http://scoledit.org/scoledition/index.php [consulté le 22/08/2020]

élèves de raconter une histoire dans laquelle il fallait introduire séparément trois phrases tout en conservant l'ordre donné. Elles ne pouvaient pas être modifiées par les élèves car elles étaient écrites sur des bandelettes en papier ou recopiées à l'identique et devaient être collées dans le texte rédigé (cf. Figure 5). Ces phrases sont les suivantes : « Elle habitait dans cette maison depuis longtemps. » ;« Il se retourna en entendant ce grand bruit. » et « Depuis cette aventure, les enfants ne sortent plus la nuit. ».

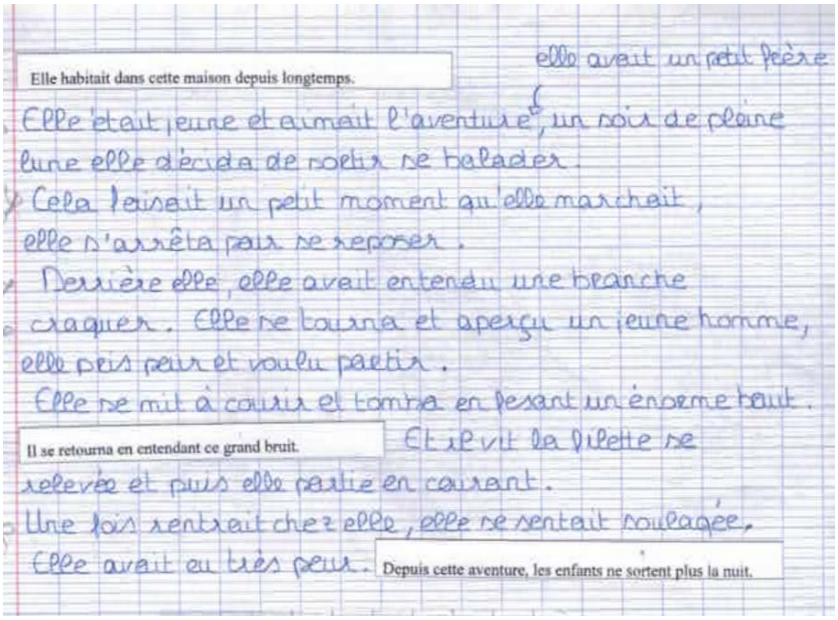

Figure 5 : Scan d'une copie d'un élève de 3ème [14]

Ainsi, pour les projets Scoledit et Resolco les données sont dites « suscitées par la recherche ». C'est-à-dire que ce sont les chercheurs qui sont allés au contact des classes et qui ont donné une consigne. A l'inverse, le corpus Ecriscol d'où proviennent mes ressources de 1ère année de licence est dit « écologique ». En effet, les données ont été produites dans le cadre de la classe, soit indépendamment de la recherche, et ont ensuite été récupérées par les chercheurs.

Si pour le projet Scoledit les données dont je dispose relèvent d'une étude longitudinale (suivi des mêmes cohortes d'une année à l'autre), il n'en est pas de même pour l'ensemble de mes ressources ; celles au-delà de l'école primaire relèvent d'un recueil de textes dans des classes en fonction des opportunités. L'étude que je mène dans le cadre de

[14] Corpus Resolco, élève R1

E-Calm relève donc d'une approche transversale plutôt que longitudinale. Le Tableau 2 est un récapitulatif de toutes les données dont je dispose pour ce travail.

Niveau	Nb textes	Nb tokens
CP	337	9632
CE1	337	24226
CE2	337	43539
CM1	337	53440
3ème	43	14450
Licence 1	23	2864

Tableau 2 : Données du corpus E-Calm utilisées durant ce travail

1.2. Pré-traitements des données

Avant que je ne les récupère, ces données ont suivi plusieurs étapes de traitement. En effet, une fois les copies collectées, il a fallu les rendre utilisables numériquement parlant.

Tout d'abord, les productions écrites des élèves ont été scannées (cf. Forme 3 dans la Figure 6), puis transcrites (cf. Forme 3 dans la Figure 6). Pour cela, les textes des élèves ont été transcrits numériquement selon une approche de type diplomatique. C'est-à-dire qu'il fallait transcrire le texte au plus près de l'original : avec ses erreurs, ses retours à la ligne, ses insertions de texte, sa ponctuation, etc. Bien entendu, il est impossible de transcrire l'intégralité des informations contenues sur une copie comme les dessins par exemple. L'ensemble du protocole de transcription est détaillé dans un guide de transcription[15] et fera l'objet ultérieurement d'une publication spécifique.

Pour permettre l'analyse des textes via des outils numériques, ils ont ensuite été normalisés (cf. Forme 3 dans la Figure 6), c'est-à-dire corrigés afin d'avoir un point de comparaison entre la production et une certaine norme attendue. Ces règles sont décrites dans un guide de normalisation[16]. Le principe de base de cette étape de normalisation est de rester le plus fidèle à ce qu'a voulu faire l'élève. Ainsi, si l'orthographe, les accords, les mauvaises segmentations, etc. sont corrigés, les erreurs de choix de temps sont conservées. Cela permet de ne pas perdre l'information sur les temps verbaux utilisés par les élèves.

[15] Disponible à l'adresse suivante : https://github.com/RachelGaubil/Outiller_description_morpho_adjectivale_primaire_universite/tree/master/guides

[16] Disponible à l'adresse suivante : https://github.com/RachelGaubil/Outiller_description_morpho_adjectivale_primaire_universite/tree/master/guides

Ces transcriptions et normalisations ont été réalisées par les chercheurs du projet ainsi que par des stagiaires et vacataires. La Figure 6 est un exemple de ces différentes étapes sur la production d'un élève de CE2[17].

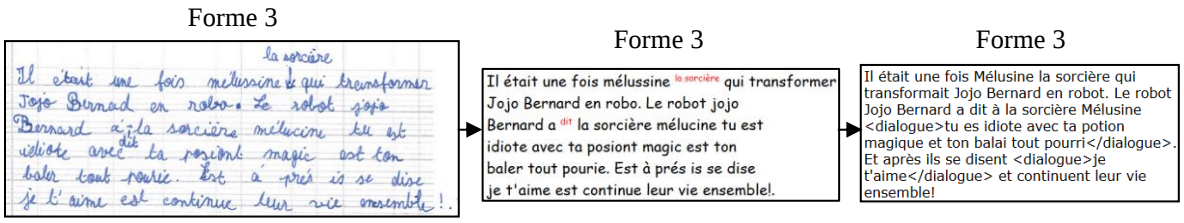

Figure 6 : Chronologie des traitements sur le texte d'un élève de CE2 pour le projet Scoledit

On remarque dans cet exemple que lors de la normalisation « la sorcière qui transformer » a été corrigé en « la sorcière qui transformait » et « après ils se dise » en « après ils se disent ». Les temps des verbes n'ont donc pas été changés pour rendre le tout plus cohérent. N'est corrigée ici que l'orthographe du mot, comme par exemple « posiont magic » qui devient « potion magique ».

Ce traitement des données a été réalisé selon les mêmes guides de transcription et de normalisation dans l'ensemble des sous-projets. Toutefois, au vu du niveau d'avancement de ces différents projets, les outils et méthodes utilisés ont pu différer. Pour constituer un corpus homogène, partageable et utilisable par les différents outils du projet, l'ensemble des transcriptions est stocké au format XML-TEI. Une première version de ce corpus est accessible sur la plateforme Ortolang[18]. Les normalisations, elles, ne sont pour le moment pas partagées et ne sont utilisées que pour la tâche 2 du projet E-Calm (tâche orthographe).

Si l'on reprend l'exemple de la chaîne de traitement précédente, derrière la Forme 3 se trouve la forme XML présentée en Figure 7.

[17] Corpus Scoledit, CE2, élève 932
[18] https://www.ortolang.fr/market/corpora/e-calm [consulté le 23/08/2020]

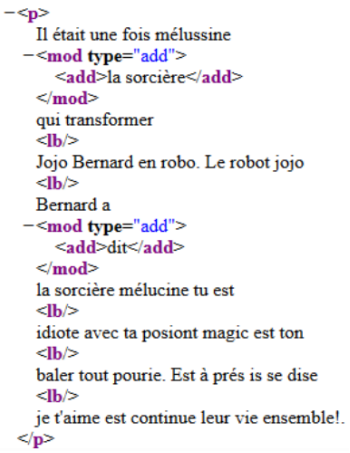

Figure 7 : Représentation XML de la production présentée en Figure 6

Une fois ces étapes terminées, les données sont passées dans un aligneur qui les segmente en tokens[19], aligne les formes produites aves les formes normées et leur ajoute des informations supplémentaires. Cet outil a été développée par Claire Wolfarth durant son doctorat (Wolfarth, 2019). Ce sont les résultats de ce système qui constituent la base de mon travail. Cet aligneur nommé *AliScol* sera présenté plus en détail dans la partie 1. Alignement des productions (Chapitre 3. Prétraitement des données, p. 35).

Maintenant que les ressources dont je dispose ont été présentées, je vais préciser les questions de recherche qui font l'objet de ce mémoire.

2. Problématique et méthodologie

Comme énoncé dans l'introduction, ce travail de mémoire s'insère dans la tâche 2 du projet E-Calm, à savoir la description des compétences en orthographe lexicale et grammaticale. La première s'intéresse à la graphie correcte d'un mot en dehors de son contexte tandis que la deuxième concerne ses transformations : les accords en genre et en nombre, les conjugaisons, etc. Cependant, au vu de l'étendue du sujet, il n'était pas possible que le projet porte sur l'ensemble des questions concernant l'orthographe. C'est pourquoi il a été choisi de se focaliser sur des zones qui semblent poser problème à tous les niveaux de

[19] Un « token » est un terme anglais qui, en linguistique, désigne une unité lexicale.

scolarité : la morphologie verbale et adjectivale ainsi que la question des lettres dérivatives (ex. le *–t* dans la forme chat).

L'ensemble de ces questions est traité par l'équipe du projet et mon travail porte plus spécifiquement sur la question des erreurs liées à la morphologie adjectivale. Ici, plusieurs questions peuvent se poser : quels sont les problèmes rencontrés par les élèves lors de l'apprentissage de l'accord de l'adjectif ? Le ratio d'erreurs sur la base du mot et sur ses flexions est-il homogène ? Cela signifierait que la difficulté est équivalente entre apprendre l'orthographe lexicale d'un mot et ses transformations en contexte.

Il a été montré dans une étude longitudinale portant sur l'acquisition de la morphologie verbale du CP au CE2 (Wolfarth *et al.*, 2018) que l'apprentissage de la base du verbe et de ses désinences se faisait à deux vitesses. En effet, au CP le taux d'erreurs portant sur les bases et sur les désinences est presque le même (49,3% et 45,5%) alors qu'en CE2 il n'y a plus que 18,3% d'erreurs sur les bases et 32,3% sur les désinences. Il y a donc une très nette progression des performances sur les bases des verbes et une moins marquée sur les désinences.

A notre connaissance, aucune étude n'a encore été menée à propos de l'acquisition des bases séparément des flexions de l'adjectif. Cependant, on sait que, comme pour les verbes, le pluriel des adjectifs est plus compliqué à maitriser que celui des noms (Fayol *et al.*, 1999, cité par Lefrançois, 2009). Grâce à ce qui a pu être observé précédemment à propos de l'acquisition des bases et désinences du verbe peut-être pourrait-on faire un parallèle avec les bases et flexions de l'adjectif. La question pourrait donc se poser de savoir si l'apprentissage de la morphologie adjectivale suit le même chemin que celui de la morphologie verbale. C'est-à-dire que les erreurs qui resteraient au fur et à mesure que l'on avance dans les niveaux scolaires seraient celles qui concernent les flexions de l'adjectif et non sa base. C'est l'hypothèse sur laquelle je m'appuierai lors de l'analyse des données.

Par ailleurs, le corpus final du projet étant vaste, il ne permet pas d'effectuer un traitement manuel. C'est pourquoi, l'objectif de ce mémoire est de réussir à outiller la description linguistique des adjectifs en production, avec comme visée plus globale d'aider les linguistes et didacticiens à comprendre où se situent les réussites et les échecs et comment ces résultats évoluent avec le temps.

Pour réaliser ceci, la méthodologie suivie s'appuie sur celle utilisée pour la morphologie verbale dans le cadre du projet, à savoir une approche par comparaison (Wolfarth *et al.*, 2018) qui est fait avec les verbes. Suite à un alignement entre transcription et normalisation, cette approche propose une comparaison entre la forme produite par l'élève et la forme normée, soit correctement orthographiée (Gourdet, 2017 ; Wolfarth, 2019). Ces comparaisons se situent à plusieurs niveaux : la graphie, la phonologie et la morphologie. Par exemple, la comparaison entre la forme produite « étais » et la forme normée « était » produit une différence au niveau graphique (-s vs -t) mais pas au niveau phonologique. Au niveau morphologique, on compare les découpages « ét+ai+s » et « ét+ai+t ». La comparaison permet de voir que l'erreur porte uniquement sur la flexion de personne mais que le radical et la flexion de temps sont corrects. Il en va de même pour les adjectifs qui seront analysés selon ces trois niveaux. Les comparaisons graphiques et phonologiques étant faites par l'outil d'alignement *AliScol*, mon travail traitera le niveau morphologique en proposant un découpage en radical et flexions puis en comparant ces découpages comme par exemple : « vert+e+s » contre « vert+ø+s ».

La suite de ce mémoire propose un état de l'art divisé en deux parties. La première s'intéresse d'un point de vue linguistique et didactique à la morphologie adjectivale et les problèmes qu'elle entraîne en apprentissage. La deuxième est consacrée aux modèles et outils TAL existants concernant la morphologie flexionnelle. Je présenterai ensuite la modélisation linguistique du comportement adjectival qui sert de base au chapitre suivant : la réalisation d'un module de traitement des adjectifs permettant l'analyse automatique des données. Après quoi, j'analyserai les résultats produits par ce système et conclurai en donnant quelques perspectives possibles.

Chapitre 2. État de l'art

Comme expliqué précédemment, ce travail vise à trouver une méthodologie permettant de décrire finement de façon automatique les erreurs et les réussites portant sur l'adjectif afin de mieux comprendre les difficultés des élèves et adapter les pédagogies. Avant de concevoir un tel outil, je vais explorer dans un premier temps ce que dit la littérature sur les difficultés des enfants autour de l'apprentissage de la morphologie adjectivale. Dans un deuxième temps, je m'intéresserai aux diverses approches TAL concernant le traitement de la morphologie.

1. Situer les difficultés de l'accord de l'adjectif

« L'adjectif varie en genre et en nombre » est une phrase que nous avons probablement tous beaucoup entendue lors de notre scolarité. Mais ce n'est peut-être pas la seule difficulté rencontrée par les élèves. Je commencerai donc par m'interroger sur l'influence de la variabilité phonique sur la graphie et m'intéresserai ensuite à la variabilité morphologique des adjectifs. Dans un troisième temps, la réflexion sera recentrée autour des marques de genre. Enfin je rechercherai d'autres sources de difficulté possibles chez les élèves.

1.1. Variabilité phonique

Les élèves se fondent sur la variabilité phonique dans leur apprentissage (Cogis & Brissaud, 2019). En effet, un changement sonore implique (souvent) un changement de graphie. Ce qui explique que les élèves ont tendance à oraliser les mots pour pouvoir les écrire. Cependant, il a été montré que la variabilité diffère de la forme orale à la forme écrite du mot. Ainsi, on trouve plus d'adjectifs invariables phoniquement que graphiquement (Séguin, 1973 ; Cogis & Brissaud, 2019), comme par exemple « calme », « joli », « essentiel », etc.

Ce constat explique les résultats observés par Cogis & Brissaud (2019) dans leur observation du marquage du féminin : 96% d'adjectifs sont correctement orthographiés si

leur variabilité morphologique est également présente à l'oral, contre 44,7% de réussite pour les adjectifs invariables phoniquement et variables graphiquement.

1.2. Variabilité morphologique

La variabilité en genre et en nombre des adjectifs est une croyance à modérer car elle ne tient pas compte de ceux qui sont épicènes[20] (Noailly, 1999). Par ailleurs, la différence entre ces deux types de variations réside dans le fait que le genre est une catégorie obligatoire tandis que le nombre est généralement un choix en discours (Cogis & Brissaud, 2019). C'est-à-dire que le genre d'un mot est établi par définition, par exemple « chat » est masculin et « voiture » est féminin. A l'inverse, le nombre peut généralement être choisi par le scripteur comme avec « le chat dort » ou « les chats dorment ».

Ces variations sont donc repérées différemment. La plupart du temps la variabilité en nombre est marquée par la lettre *-s* (Noailly, 1999). À l'inverse le genre ne porte pas forcément de marque (Cogis & Brissaud, 2019). Or cette absence crée un « conflit cognitif » (Cogis, 2003, p.106) qui pousse les élèves à inventer une marque. En effet, on observe l'ajout de *-t* ou de *-s* (choix fait parmi les graphèmes polyvalents) en fin de mot (ex. « pointut ») par les élèves pour le masculin puisque le *-e* est souvent associé au féminin (Cogis, 2003).

1.3. Marques de genre

1.3.1. Le -e féminin ?

Le *-e* devient rapidement la représentation du féminin chez les élèves (Cogis & Brissaud, 2019). Cependant, plusieurs chercheuses s'accordent à dire que cette lettre ne peut pas être la marque du féminin. En effet, le *-e* est également très présent à la fin des verbes (Cogis & Brissaud, 2019) et des mots finissant à l'oral par une consonne comme « célèbre », « leste », etc. (Cogis, 2003). En outre, on le trouve également dans tous les adjectifs épicènes (Noailly, 1999).

Malgré ces arguments linguistiques, cette lettre *-e*, bien souvent vue comme une « lettre de fille » (Cogis, 2003 ; Cogis & Brissaud, 2019), incarne la marque du féminin (Cogis, 2003) et empêche alors sa présence à la fin d'un mot masculin comme dans

[20] C'est-à-dire, les adjectifs dont la forme ne varie pas lorsque le genre change comme « habile », « pauvre », etc.

« terribl ». On observe donc une confusion entre le genre grammatical et l'opposition de sexe (Cogis & Brissaud, 2019).

1.3.2. Entre genre grammatical et genre humain

Ainsi, il est courant de voir dans les copies écrites par des filles des *-e* à la fin de chaque adjectif. Cogis & Brissaud (2019) ont demandé à des élèves des justifications sur leur façon d'écrire. C'est ainsi qu'une élève de CE2 expliquait « sérée » dans la phrase « j'avais pris du 34 mais c'était trop sérée sur mes pieds » par le fait que « c'était à moi » (Cogis & Brissaud, 2019, p. 59). On observe donc que le genre du narrateur ou du scripteur a une influence sur la façon dont les élèves accordent les mots.

Ce type de malentendu a également lieu lorsque le genre du locuteur est différent du genre narratif. Le *-e* étant pour les filles, il arrive que les garçons ne veuillent pas l'employer dans une histoire qu'ils écrivent même si leur personnage est féminin ; réciproquement, cela choque certaines filles que des garçons utilisent le *-e* à la fin des mots dans leurs textes narratifs (Cogis, 2003). Le respect de la norme orthographique repose donc sur la représentation du genre que possède l'élève (Cogis, 2003). Par ailleurs, certaines chercheuses se sont demandé s'il existait également d'autres obstacles à cet apprentissage des accords.

1.4. Autres sources de difficultés

1.4.1. Valeur sémantique et structure syntaxique

Cogis & Brissaud (2019) se sont intéressées à l'impact possible de la valeur sémantique[21] et de la structure syntaxique[22] sur la complexité du respect des accords adjectivaux. Elles ont observé les productions écrites d'élèves en relevant les pourcentages d'accords corrects dans différents cas de figure proposés[23].

Cette expérimentation a montré qu'il n'y avait pas de différence significative de réussite lorsque l'adjectif était en position d'épithète ou d'attribut, suggérant que la structure syntaxique n'intervient pas. A l'inverse, les chercheuses ont pu constater que les adjectifs

[21] Testée ici par l'opposition entre les adjectifs qualifiants et classifiants.
[22] Testée ici par l'opposition entre les adjectifs épithètes et attributs.
[23] L'expérimentation a été menée en 2009 avec 247 élèves de CM2 provenant de villes différentes. 12 items ont été utilisés : 8 adjectifs différents, dont 4 dictés deux fois dans des contextes différents. Ces items étaient répartis de la façon suivante :
- 6 adjectifs variables à l'oral et 6 invariables à l'oral,
- 6 adjectifs qualifiants et 6 classifiants,
- 8 adjectifs épithètes et 4 attributs.

classifiants (ex. « national ») posaient plus de difficultés que les adjectifs qualifiants (ex. « petit »). Cependant, il est difficile d'attribuer cette différence à la valeur sémantique des adjectifs car le type de finales (consonantique ou vocalique) pourrait également entrer en jeu. Par ailleurs, les terminaisons en *-é* des participes passés utilisés comme adjectifs peuvent également faire émerger des difficultés.

1.4.2. Participes passés

Cogis & Brissaud (2019) expliquent qu'il est très courant de trouver des participes passés en position d'adjectifs et qu'ils sont bien souvent mal orthographiés : les élèves utilisent notamment *-er* comme terminaison. L'erreur s'explique par le fait que le participe passé appartient à la fois au système nominal et au système verbal ; or *-er* est la marque d'identité des verbes. On passe ici d'un problème de morphologie à un problème de morphosyntaxe.

Ainsi, les obstacles ne manquent pas pour arriver à la bonne graphie d'un adjectif et la plupart des erreurs sont dues aux variations morphologiques de l'adjectif. Dans l'optique de poursuivre l'observation de ces problèmes récurrents, la partie suivante s'intéresse aux différentes méthodes d'analyse morphologique automatique.

2. Le TAL et l'analyse morphologique

« En traitement automatique, l'analyse morphologique consiste à segmenter un texte en unités élémentaires auxquelles sont attachées des connaissances dans le système : une fois cette segmentation effectuée, ce n'est plus le texte qui est manipulé, mais une liste ordonnée de telles unités (ou plusieurs listes, en cas d'ambiguïtés, réelles ou « artificielles »). » (Fuchs *et al.*, 1993). L'analyse morphologique est donc la première étape d'une analyse automatique de texte et c'est pourquoi il est primordial qu'elle soit correctement exécutée car les autres niveaux s'appuieront sur ses résultats.

Il existe plusieurs approches pour réaliser une analyse morphologique en traitement automatique des langues. Les trois principales, présentées par Blanchard (2006) sont les suivantes : les dictionnaires de formes fléchies, les automates à états finis et le modèle à deux niveaux.

2.1. Dictionnaires de formes fléchies

Très utilisés dans les systèmes de correcteurs orthographiques, les dictionnaires de formes fléchies sont, comme leur nom l'indique, une liste des formes fléchies[24]. Par exemple, l'analyseur INTEX se sert du dictionnaire de formes fléchies DELAF contenant plus de 800 000 entrées (Silberztein, 1996). Celles-ci ressemblent aux lignes suivantes :

avions, avoir.V :I1p, avion,.N:mp
est,être.V :P3s

Ici, « est » correspond au verbe (V) « être » conjugué à la première personne du singulier au présent (P3s) et « avions » est soit l'imparfait à la première personne du pluriel (I1p) du verbe (V) « avoir » soit le masculin pluriel du nom « avion » (N:mp). Ainsi, chaque entrée du dictionnaire associe une forme fléchie à son lemme, à sa partie du discours et à ses informations flexionnelles. Cela permet un faible traitement informatique puisqu'il suffit de découper le texte en mot et de chercher ses derniers dans le dictionnaire des formes fléchies.

Cependant, cette approche ne permettant pas d'accéder à un découpage en *radical – flexions*, il sera impossible d'analyser des formes erronées puisqu'elles provoqueront une erreur de la part du système. Ainsi, la forme au féminin singulier « nouvele » sera simplement étiquetée comme forme inconnue. Cette approche ne convient donc pas aux objectifs de ce projet qui aimerait proposer une analyse plus fine du type : erreur sur le radical (« nouvel » vs « nouvell ») mais marquage en genre correct. C'est pourquoi il faut maintenant s'intéresser aux automates à états finis qui, eux, permettent un découpage des mots.

2.2. Automates à états finis

En effet, les automates à états finis, séparent les mots en *base – affixe(s)*. C'est-à-dire qu'ils décomposent le mot présenté en cherchant toutes les combinaisons possibles. Par exemple le mot « jolis » peut être découpé de plusieurs façons :

j+olis *jo+lis* *jol+is* *joli+s*

C'est le dernier découpage qui sera accepté puisqu'il correspond à une combinaison *base - affixe* contenue dans le dictionnaire. Cependant, si le système rencontre une flexion

[24] Les formes fléchies sont les formes dérivées d'un lemme. Ce dernier est la forme canonique d'un mot, soit le masculin singulier ou un verbe à l'infinitif (https://dictionnaire.lerobert.com/definition/lemme [consulté le 02/09/2020]). Par exemple, « infinies » et « infiniment » sont des formes fléchies du lemme « infini ».

qui modifie la base comme « neuf – neuve » il devra pouvoir analyser « neuf », « neuf+s », « neu+ve » et « neu+ve+s » et donc avoir dans son lexique une base sans réelle justification linguistique (« neu »). Cela se voit bien dans la représentation suivante (cf. Figure 8) où « e0 » correspond à l'état initial et « e2 », « e3 », « e4 » et « e5 » aux états finaux.

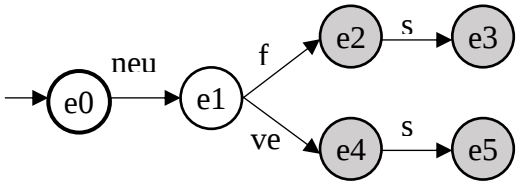

Figure 8 : Automate à états finis représentant les formes fléchies possibles du lemme « neuf ».

De plus, cela augmentera fortement le nombre de données (bases, flexions) du système. Par ailleurs, il y a ici un autre problème, commun avec la méthode du dictionnaire des formes fléchies : l'analyse de formes erronées. En effet, si la forme présentée à l'automate est « neufe » au lieu de « neuve » l'automate n'arrivera pas à un état final et cela engendrera une erreur. Le système ne sera donc pas capable de renvoyer un découpage tel que « neuf + e » puisque cette combinaison n'existe pas. Cette méthode n'est donc pas non plus adaptée à notre projet.

Vient enfin le modèle à deux niveaux qui consiste à établir « une correspondance entre une représentation lexicale, sous la forme d'une séquence de morphèmes[25] et de quelques symboles auxiliaires tels que frontière de morphème, frontière de mots, et la représentation orthographique du mot » (Wehrli, 1997). C'est cette approche qui sera présentée dans la partie qui suit.

2.3. L'Analyse morphologique à deux niveaux

Je m'attacherai ici à décrire l'approche computationnelle à deux niveaux, présentée par Fradin (1994) ; celle-ci est constituée de lexiques et de règles. La particularité de cette approche est qu'elle se base sur les lexèmes[26] et non sur les morphèmes (qui sont habituellement l'unité choisie en analyse morphologique à deux niveaux).

[25] Le morphème est une unité minimale de signification (https://www.cnrtl.fr/lexicographie/morph%C3%A8me [consulté le 16/01/2020]).

[26] Un lexème est un morphème lexical, soit l'unité minimale de signification appartenant au lexique (https://www.cnrtl.fr/definition/lex%C3%A8me [consulté le 16/01/2020]) ; tandis qu'un morphème est la plus petite unité de sens et n'est pas forcément lexical.

2.3.1. Lexiques

Le lexique se décompose en un « lexique-base » et plusieurs « lexiques continuateurs » (lexique des suffixes adjectivaux, lexique des suffixes nominaux, etc.) (Fradin, 1994, p.30). Le lexique-base, comme son nom l'indique, contient la base des mots ; tandis que les autres lexiques contiennent toutes les formes réalisables.

Tous ces lexiques sont représentés sous la forme d'arbres de lettres dont les derniers nœuds sont les catégories syntaxiques. Si ces dernières ne sont pas indiquées : on trouve alors des informations morphologiques ou des instructions indiquant à quel lexique continuateur se référer pour rallonger le lexème.

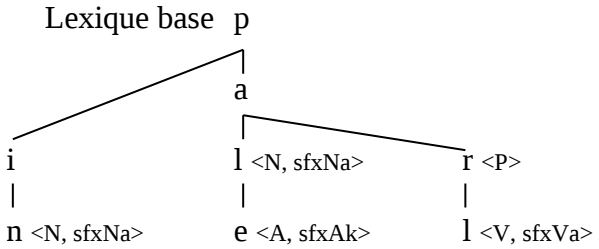

Figure 9 : Exemple d'une partie du lexique base (Fradin, 1994, p.30)

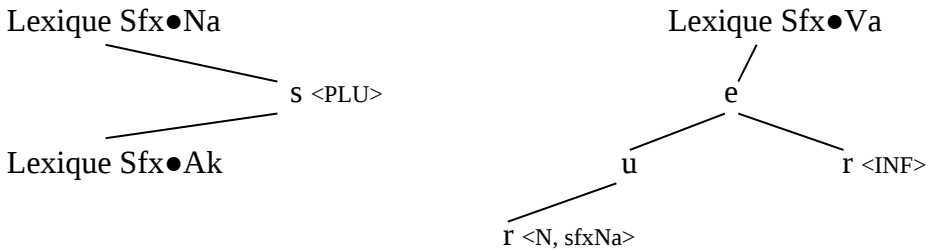

Figure 10 : Deux exemples de lexiques continuateurs de suffixes (le premier concerne les nominaux de type a et les adjectifs de type k ; le deuxième concerne les verbes de type a) (Fradin, 1994, p.30)

Par exemple, le lexique base de la Figure 9 contient les lexèmes *pin*, *pal*, *pâle*, *par*, *parl* ; et il peut être complété grâce aux deux lexiques continuateurs de la Figure 10 qui ajoutent des suffixes. Cela donne les mots : *pins, pals, pâles, parler, parleur*.

En plus de ces lexiques, l'approche computationnelle à deux niveaux nécessite différents types de règles.

2.3.2. Règles morphologiques et phonologiques

Ces règles ont pour but d'autoriser ou non la mise en relation entre deux suites de symboles : une suite de surface et une suite lexicale. La première correspond à la forme

produite du mot (la forme à analyser) ; tandis que la seconde est celle contenue dans les lexiques. La particularité de ces règles réside donc dans le fait qu'elles traitent deux symboles à la fois : un symbole de surface et un du lexique.

Ce double traitement est possible grâce aux transducteurs à états finis qui représentent les règles. En effet, ils peuvent analyser deux symboles à la fois si chacun est écrit sur une bande différente.

En plus des règles morphologiques, cette approche dispose également de règles phonologiques, elles aussi représentées par des transducteurs à états finis. Les règles phonologiques s'appliquent habituellement les unes à la suite des autres (Kay & Kaplan, 1983, cités par Fradin, 1994). Cette architecture en cascade n'étant pas la plus optimale en reconnaissance (un phonème de surface peut correspondre à plusieurs phonèmes du lexique), Koskenniemi (1983, cité par Fradin, 1994) a proposé un traitement parallèle des transducteurs.

La compilation des règles par des transducteurs à états finis empêche le ralentissement du temps de traitement qui peut avoir lieu lorsque la complexité augmente. Par ailleurs, cette approche peut aussi bien être employée en analyse qu'en génération.

Cependant, cette approche présente le même défaut que les précédentes concernant les formes erronées : elle ne peut pas les analyser. C'est pourquoi je vais maintenant me tourner vers une approche d'analyse morphologique flexionnelle qui se base sur un découpage linguistique des formes. En effet, cette approche semble a priori intéressante dans le cadre de notre projet puisqu'elle permet de trouver une interprétation aux bases et aux flexions extraites.

2.4. Le modèle du CRISS

Dans cette partie je présenterai le modèle d'analyse morphologique flexionnelle proposé au sein du laboratoire CRISS[27] par Lallich-Bodin (1987). Il se rapproche du modèle à deux niveaux par ses règles de régularisation (cf. 2.4.1.2 Fonction calculatoire, p. 28) mais il est plus intéressant dans le cadre de ce projet car il propose un découpage des mots en *radical - flexions* linguistiquement justifié. C'est-à-dire qu'il découpe les termes de façon à en interpréter chaque partie, par exemple « neuve » sera découpée en « neuf* + e + ø »

[27] Centre de Recherche en Informatique appliquée aux Sciences Sociales (ancien centre appartenant à l'université Pierre Mendès-France).

(« f* » correspond à la lettre « v », cf. Annexe 3, p. 92). Ici, la base « neuf* » a un sens connu, contrairement à « neu » dans la représentation précédente (2.2 Automates à états finis, p. 24). De plus « e » et « ø » indiquent respectivement une marque de genre et une absence de marque de nombre. De ce découpage il est donc possible de déduire que « neuve » est la forme au féminin singulier de « neuf ».

Je vais commencer par présenter les différentes fonctions de ce modèle pour expliquer ensuite sa chaîne de traitement. Cependant, j'ai fait le choix de ne détailler que les parties concernant l'adjectif puisque c'est le sujet principal de ce travail.

2.4.1. Principe

L'analyseur morphologique décrit par Lallich-Boidin (1987) possède deux fonctions : une classificatoire qui fait correspondre une catégorie morphologique à une forme et une calculatoire qui calcule la base d'une forme. Notons que la base correspond à la racine d'un mot et la forme à une de ses formes dérivées.

2.4.1.1. **Fonction classificatoire**

Lors de l'étape de classification on associe à chaque base les informations qu'il est nécessaire de connaître pour l'analyse des mots. Dans cette organisation, l'adjectif fait partie du groupe des *nominaux* avec les substantifs ; ce qui permet alors de distinguer l'adjectif du nom est la sous-catégorie syntaxique *NA* qui attribue en fonction du contexte la valeur *nom*, *adj* ou *nan*[28]. Les deux sous-catégorisations flexionnelles qui peuvent concerner l'adjectif sont *GR* et *NB* qui indiquent respectivement le genre et le nombre.

2.4.1.2. **Fonction calculatoire**

Ce modèle divise les flexions en deux types : les flexions du *nom-adjectif* (genre et nombre) et les flexions du *verbe* (temps et personne). L'ordre d'apparition des flexions décrit dans la Figure 11 se lit de droite à gauche.

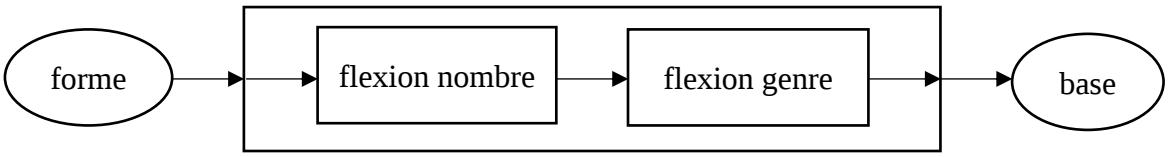

Figure 11 : Extrait du schéma d'ordre d'apparition des flexions concernant l'adjectif (Lallich-Boidin, 1987, p. 8).

[28] La valeur *nan* est donnée dans le cas où le contexte ne permet pas de discerner si le mot est un substantif ou un adjectif.

Dans le cas général, on part donc de la racine à laquelle on ajoute la marque de genre, ensuite celle du nombre pour enfin arriver à la forme dérivée voulue. Par la suite, afin de décrire les différentes variations de ce comportement, un certain nombre de modèles flexionnels[29] ont été élaborés. Ils sont de la forme suivante :

	nom ou adjectif invariable au masculin, et prenant *un* *e au féminin			
gros	**[mas, nbn]**	**[*e]**	**[]**	**fem, sng**
		[*e]	**[s]**	**fem, plu**
	nom ou adjectif dont le féminin se forme par ajout de *te*			
rigolo	**[mas, sng]**	**[]**	**[s]**	**plu**
		[te]	**[]**	**fem**
		[te]	**[s]**	**fem, plu**

Figure 12 : Extrait des modèles de flexions des nominaux (Lallich-Boidin, 1987, p. 9-10).

Ces modèles permettent pour l'instant uniquement de décrire les cas généraux, l'objectif étant d'y ramener les cas particuliers afin de diminuer le nombre de modèles. Pour cela, il faut opérer des régularisations sur les mots que l'on ne peut pas analyser grâce à ces modèles. Il y a deux types de régularisations possibles : sur la forme ou sur la base. A noter que ces régularisations sont comparables aux règles de la morphologie flexionnelle à deux niveaux vue précédemment.

2.4.1.2.1. Régularisations de base

Les régularisations sur la base du mot concernent les termes dont la racine est influencée par la flexion et n'est donc pas présente dans le dictionnaire, par exemple : « *heureus+e+s => heureux+e+s* » (Lallich-Boidin, 1987, p. 11). Pour régulariser ces comportements, les substitutions suivantes seront appliquées :

'T en T	*inquiet*	modèle *avocat*
'C en C	*sec*	modèle *bon*
'F en F	*bref*	modèle *bon*
'R en R	*léger*	modèle *avocat*
S en 'S	*exprès*	modèle *gros*
^C en S	*frais*	modèle *gros*
GN en N	*malin*	modèle *avocat*
C en X	*doux*	modèle *anglais*
S en X	*heureux*	modèle *anglais*
S en X	*faux*	modèle *gros*

[29] Le symbole « * » dans la Figure 12 désigne un dédoublement de consonnes. Il peut parfois signifier d'autres choses telles que l'ajout d'une lettre, cela est expliqué plus précisément en Annexe 3 (p. 92).

Figure 13 : Liste des substitutions sur les bases (Lallich-Boidin, 1987, p. 12)

Ainsi, la forme adjectivale « *légère* » est régularisée en « *légere* » puis est analysée en « *léger+e+ø* ». Le modèle de comportement est donc ramené au modèle général des adjectifs prenant un –*e* au féminin et un –*s* au pluriel (modèle avocat).

2.4.1.2.2. Régularisations de forme

En effet, si malgré les régularisations de base l'analyse se termine par un échec, on effectue alors une régularisation sur la forme. L'idée qui prédomine dans cette démarche est que le -*s* est la marque du pluriel en français ; or certains pluriels comme *travaux* se terminent par un -*x* et ne correspondent donc pas au cas général. Pour traiter ce type de cas il y a deux possibilités :

- Il faut faire un découpage du mot de la façon suivante : *trava + il* et *trava + ux*. Cependant *trava* ne signifie rien et n'est pas dans le dictionnaire utilisé.
- Il faut analyser directement la forme *travaux*. Pour cela, il faudrait substituer -*ux* par -*ils* (cas général de la formation du pluriel en –s) puis recommencer l'analyse avec la forme *travails*. Si le mot *travail* est bien enregistré comme suivant le modèle du mot *coup* (cf. Figure 12) alors le découpage sera valide et produira : *travaux : travail [mas, plu]*.

La seconde solution est celle qui a été choisie. Par ailleurs, il y a plusieurs substitutions de la sorte à effectuer ; il faut donc les ordonner afin de ne pas produire d'analyses parasites. Les substitutions sont ordonnées de la manière suivante :

1.	AUX en AILS	*travail*	modèle *coup*
2.	AUX en ALS	*journal*	modèle *coup*
3.	UX en US	*eau*	modèle *voiture*
		chapeau	modèle *coup*

Figure 14 : Liste des substitutions sur les formes (Lallich-Boidin, 1987, p. 11)

Il est important de se rappeler que toutes ses modifications ne sont que temporaires. En effet, elles servent uniquement à diminuer le nombre de modèles et de flexions nécessaires pour l'analyse morphologique.

2.4.2. Chaîne de traitement

2.4.2.1. **Stratégies d'analyse**

La première stratégie d'analyse consisterait à mettre toutes les formes possibles dans un dictionnaire. La mise en œuvre est simple, cependant cela pose un souci dû à la taille du dictionnaire et au nombre important de redondances. En effet, toutes les informations d'un mot sont également indiquées au niveau de chacune de ses formes dérivées.

La seconde stratégie consisterait donc à ne mettre dans le dictionnaire que les informations que l'on ne peut pas calculer. Le dictionnaire ne contiendrait donc qu'une seule entrée pour chaque mot invariable, nom et adjectif : leur base. Pour les verbes, le dictionnaire aurait autant d'entrées que de bases différentes pour chaque verbe (1 à 7).

C'est cette dernière stratégie qui a été choisie. Cependant, la démarche de l'analyseur peut tout de même être très coûteuse car il cherche tous les découpages réalisables. Or, aucun découpage n'est possible pour les mots invariables. C'est pour cela que la chercheuse a décidé d'ajouter dans le dictionnaire une variable booléenne *STOP* qui sera égale à *1* si le mot analysé est invariable. C'est donc le dictionnaire qui guidera l'analyseur.

2.4.2.2. **Données**

Ainsi, cette méthode d'analyse morphologique nécessite quatre types de données : un dictionnaire, les listes des flexions, les listes des compatibilités de flexions ainsi que l'ensemble des modèles linguistiques. Le dictionnaire contient :

- Les bases prétraitées[30] : le mot entier s'il s'agit d'un invariable, sinon son lemme.
- L'entrée lexicale (correspondance de la base prétraitée avec la représentation habituelle du mot, par exemple : *c*ant* → *chant*).
- Les informations expliquées dans la partie 2.4.1.1 Fonction classificatoire (p. 28).
- Le booléen *STOP* qui est égal à *1* si le mot est invariable.
- Le booléen *SYNT* qui est égal à 1 si le schéma rectionnel[31] du mot est indiqué.

Les modèles linguistiques sont les mêmes que ceux définis dans 2.4.1.2 Fonction calculatoire (p. 28) avec une simple différence d'expression :

[30] Avant toute chose, un prétraitement graphique a été effectué afin de régulariser le texte (cf. Annexe 3, p. 87).
[31] Schéma indiquant de quelle façon le mot peut admettre un ou des compléments (si cela est possible).

coup *+ [ø] + [ø] (mas, sng) ;* ***coup*** *+ [ø] + [s] (mas, plu) ;*

Figure 15 : Exemple de modèle linguistique (Lallich-Boidin, 1987, p. 23)

Une fois toutes ces données créées et organisées, il ne reste plus qu'à lancer l'analyseur sur le mot souhaité.

2.4.2.3. **Algorithme**

Cet analyseur fonctionne avec deux modules : un spécifique aux verbes (module *analyse-verbe*) et un autre pour le reste des mots (module *analyse-nom*). Les régularisations de base ont lieu dans le module concerné. Si malgré les régularisations de base, l'analyse échoue toujours alors on effectue la régularisation de forme et l'on relance les deux modules sur la nouvelle forme obtenue.

3. Comparaison des systèmes d'analyse morphologique

En ce qui concerne les dictionnaires de formes fléchies et les automates à états finis, ils ont été rapidement écartés de la liste des possibilités. En effet, ils ne permettent ni l'un ni l'autre un découpage en *base-flexion* et engendrent une erreur à la moindre forme erronée.

Les deux dernières approches présentées permettent un découpage et possèdent chacune des avantages et des inconvénients.

Le système présenté par Lallich-Boidin (1987) ne traite que les mots dont les formes dérivées sont calculables à partir de leur racine. Cependant, le découpage est intéressant car il est adapté aux besoins du projet : il découpe les mots en *base – flexions* et même plus précisément en *base – flexion de genre – flexion de nombre.* De plus, il fonctionne à partir d'un nombre réduit de modèles de flexions en privilégiant le calcul grâce aux règles de régularisation. Il diminue ainsi la quantité de données qu'il doit posséder pour effectuer le traitement demandé. Par ailleurs, il découpera de façon intéressante la forme erronée « nouvele » en « nouvel+e+ø ».

A l'inverse, l'approche de Fradin (1994) contient beaucoup de données notamment de nombreux lexiques. L'intérêt de cette approche est sa double composante avec les règles morphologiques et phonologiques. Elle serait donc la méthode la plus appropriée si nous avions également à travailler sur la phonologie des mots.

Cependant, je ne m'intéresse ici qu'à la morphologie des adjectifs et à leur découpage en *racine-flexions*. Au regard de ceci, l'approche du CRISS paraît la plus indiquée. En effet, elle permet un découpage des flexions assez fin et clair et ne nécessite pas une grande capacité de stockage.

En conclusion, l'accord de l'adjectif n'est pas forcément aisé à acquérir à cause de ses variations morphologiques qui ne sont pas toujours identifiables lorsque le mot est oralisé. Les enfants se basent donc sur les règles générales qu'ils ont apprises ou déduites afin d'orthographier les adjectifs du mieux possible. Afin d'analyser les erreurs commises pour en relever des problèmes récurrents, il serait intéressant de pouvoir utiliser un découpage morphologique automatique qui séparerait les mots en *racine – flexions*. En effet, compte tenu de l'hypothèse de départ, il y aurait une différence dans l'acquisition de la base d'un mot et de ses flexions. Il serait donc intéressant d'analyser les deux séparément. Pour cela, l'analyseur décrit par Lallich-Boidin (1987) semble être un bon point de départ car, en se basant sur un découpage linguistique, il permet de donner une interprétation aux bases et flexions obtenues.

Dans la suite de ce travail, je m'appliquerai à élaborer un prototype de système permettant d'assister l'analyse des erreurs de l'accord adjectival.

Partie 2

-

Modélisation

Chapitre 3. Prétraitement des données

L'objectif de mon projet est de pouvoir repérer où se situent les difficultés pour les élèves dans leur apprentissage de la morphologie adjectivale. Pour pouvoir observer cela, je vais donc extraire les adjectifs des productions et tenter de proposer un découpage interprétable selon l'approche par comparaison adoptée dans E-Calm.

Cependant, les productions ne sont pas directement traitables informatiquement, il va donc falloir effectuer quelques traitements initiaux dessus avant de pouvoir les utiliser. Il se trouve que les données ont déjà été transcrites puis normalisées (cf. Figure 6). Pour traiter les formes adjectivales uniquement, je m'appuie sur les résultats de l'outil *AliScol* (Wolfarth, 2019) qui aligne chaque forme produite à sa forme normée.

1. Alignement des productions

AliScol est basé sur un algorithme d'alignement dont le schéma général est le suivant :

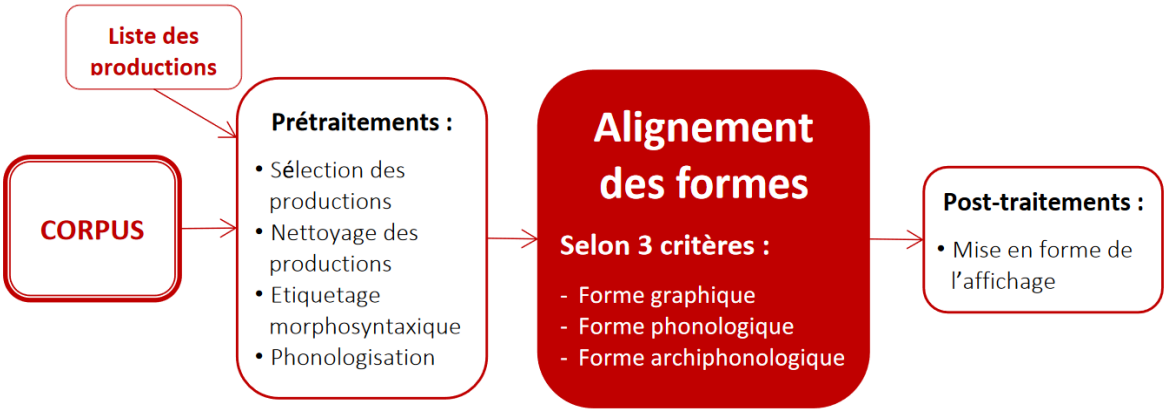

Figure 16 : Schéma général de l'algorithme d'alignement (Wolfarth, 2019, p. 158)

Dans les prétraitements qui ont dû être réalisés, deux outils ont été utilisés : *TreeTagger*[32], un étiqueteur morphologique et *LIA-PHON*[33] qui permet d'associer à une forme graphique une représentation phonologique.

[32] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger [consulté le 24/04/2020]
[33] http://pageperso.lif.univ-mrs.fr/~frederic.bechet/download.html [consulté le 24/04/2020]

AliScol sert à aligner des productions transcrites et des productions normées au niveau des formes. Cependant, la notion de forme ne convient pas en cas d'hyper ou d'hypo-segmentation. En effet, par exemple, des termes comme « lattrape » sont constitués de deux formes « l' » et « attrape ». Dans sa thèse, Wolfarth (2019) préfère ainsi parler d'alignement de segments. Cet alignement se fait à l'aide de trois critères : la graphie, la phonologie et l'archiphonologie qui ont chacun un rôle précis :

- La comparaison de la graphie permet de repérer si l'orthographe est correcte.
- L'analyse de la phonologie indique les erreurs faites qui respectent (ou presque) la phonologie attendue.
- Fondée sur les travaux de N. Catach (1995), l'archiphonologie permet de neutraliser certaines différences phonologiques telles que /o/ et /ɔ/ qui peuvent être simplement dues à des variations interlocuteurs (différence régionale ou sociale) ou à des différences de contexte d'apparition du graphème[34].

Ce sont les données obtenues (cf. Annexe 4, p. 92) une fois l'alignement lancé qui pourront être utilisées dans la suite de ce travail. Elles se présentent sous la forme d'un fichier csv au format présenté dans le Tableau 3.

Id Prod	Id Eleve	Id Classe	Niv	Long Prod	IdTok Norme	Lemme	Catégorie	Seg Norme	Seg Trans	IdTok Trans	IdSeg Trans	Statut Erreur	Statut Segm
2000_CE1	2000		CE1	143	13	il	PRO:PER	Il	Il	13	1	01-Normé	01-Normé
2000_CE1	2000		CE1	143	14	ne	ADV	n'	naver	14	1	03-Archi	03-HypoSeg
2000_CE1	2000		CE1	143	15	avoir	VER:impf	avait	naver	14	2	03-Archi	03-HypoSeg
2000_CE1	2000		CE1	143	16	pas	ADV	pas	pas	15	1	01-Normé	01-Normé
2000_CE1	2000		CE1	143	17	de	PRP	de	de	16	1	01-Normé	01-Normé
2000_CE1	2000		CE1	143	18	pouvoir	NOM	pouvoir	pouvoir	17	1	01-Normé	01-Normé

Tableau 3 : Exemple de sortie d'AliScol pour une production d'un élève de CE1

2. Enrichissement des résultat d'AliScol

AliScol apporte ainsi beaucoup d'informations utiles comme la catégorie du mot et sa forme normée par exemple. Cependant il ne possède pas toutes les caractéristiques qui

[34] Un graphème est un ensemble minimal de lettres transcrivant un phonème ou ayant une fonction morphologique ou étymologique (https://www.cnrtl.fr/definition/graph%C3%A8me) [consulté le 22/07/2020].

sont nécessaires pour la suite ; il manque les informations de genre et de nombre par exemple. En effet, il est important de savoir si la forme analysée devrait être au masculin ou au féminin, au singulier ou au pluriel afin de déterminer la découpe. Par exemple, sur un mot féminin, généralement, il faut chercher le « e » final. De même, pour les verbes les informations de mode, de temps et de personne sont essentielles pour un découpage en *radical - désinences*.

J'ai donc conçu un module d'enrichissement[35] qui récupère les fichiers produits par l'aligneur et leur ajoute trois colonnes : *genre, nombre, infover* (cf. Annexe 5, p. 94). La dernière catégorie ne concerne que les verbes et regroupe les informations de temps, de mode et de personne. Elles sont indiquées de la façon suivante :

- Genre : « m » ou « f » pour le masculin ou le féminin, ou « _ » si le mot est épicène ;
- Nombre : « s » ou « p » pour le singulier ou le pluriel, ou « _ » si le mot est invariable ;
- Informations pour les verbes : « mode:temps:personne ». Par exemple, « pratiquait » est indiqué comme « ind:imp:3s » pour l'indicatif, l'imparfait et la troisième personne du singulier. Si le verbe est à l'infinitif seul « inf » est indiqué. S'il y a ambiguïté, toutes les possibilités sont écrites séparées par un point-virgule, par exemple le mot « allons » peut être à l'impératif ou à l'indicatif présent, on écrit donc : « imp:pre:1p;ind:pre:1p ». Cette représentation suit celle du Lexique 3.83.

Par ailleurs, avant de choisir ce lexique comme référence il a fallu se renseigner sur les différentes ressources disponibles. De nos jours, sont accessibles en ligne beaucoup de dictionnaires et lexiques gratuits. Cependant des sites internet tels que Wikipedia[36] ne sont pas pratiques dans ce travail car il faudrait faire une étape d'extraction des données du site et de nettoyage pour ne conserver que ce qui est pertinent ici. Pour ce qui est des lexiques gratuits et téléchargeables, le choix s'est porté entre la liste de mots commun d'ABU[37] (1) et le Lexique 3.83[38] (2). Les informations sont représentées de la façon suivante :

(1) mot lemme Cat:Genre+NOMBRE ou Ver:ModeTemps+NOMBRE+PERSONNE

[35] Disponible à l'adresse suivante : https://github.com/RachelGaubil/Outiller_description_morpho_adjectivale_primaire_universite/tree/master/module_enrichissement
[36] https://www.wikipedia.org/ [le 10/03/2020]
[37] http://abu.cnam.fr/DICO/mots-communs.html [le 10/03/2020].
[38] Récupéré sur http://www.lexique.org/ [le 10/03/2020].

(2) mot phonétique lemme cat genre nombre … mode:temps:personne

Les « … » représentent des informations qui ne sont pas utiles dans ce contexte. Les regroupements effectués par ABU au niveau de la catégorie, du genre et du nombre ne sont pas pratiques car ils nécessitent un découpage afin d'extraire séparément les informations recherchées. C'est pourquoi je me suis basée sur le Lexique 3.83.

Le système créé parcourt donc les données extraites de l'aligneur et les complète à l'aide des informations trouvées dans ce lexique. Plus précisément, il récupère comme informations de l'aligneur la forme normée, le lemme et la catégorie. Il cherche ensuite la forme normée dans le lexique et compare le lemme et la catégorie trouvés dans ce dernier à ceux de l'aligneur. Si ce sont les mêmes, le système récupère les informations des colonnes « genre », « nombre » et « infover » du lexique et les ajoute aux les données.

Ce dernier a été modifié au fur et à mesure de l'avancement du travail car il comportait quelques incohérences qui ont engendré des problèmes lors de nos différents traitements. Le Tableau 4 montre quelques exemples des modifications qui ont dû être effectuées dans le Lexique 3.83.

	Lexique originel			**Lexique corrigé**	
Forme	**Genre**	**Nombre**	**Problème**	**Genre**	**Nombre**
furax	m	_	Le mot est invariable en genre.	_	_
animaux	m	_	Cette forme est celle du pluriel.	m	p
orange	_	_	Le mot est variable en nombre.	_	s

Tableau 4 : Exemples des modifications faites au Lexique 3.83

Certains termes ont également dû être ajoutés : par exemple, le mot « horrifique » était bien présent mais pas son pluriel « horrifiques ».

Le résultat des traitements présentés ci-dessus se présente sous la forme d'un fichier csv décrivant les données qui serviront d'entrée à notre module sur la morphologie adjectivale. Ce fichier garde le même format que celui présenté dans le Tableau 3 (p. 36) à la différence des trois dernières colonnes ajoutées. Le Tableau 5 est un exemple de sortie du module d'enrichissement dont les colonnes ne présentant pas d'intérêt dans l'explication de cette étape ont été supprimées.

Lemme	Catégorie	Seg Norme	Seg Trans	Statut Erreur	Statut Segm	Genre	Nombre	Infover
il	PRO:PER	Il	Il	01-Normé	01-Normé	m	s	_
ne	ADV	n'	naver	03-Archi	03-HypoSeg	_	_	_
avoir	VER:impf	avait	naver	03-Archi	03-HypoSeg	_	_	ind:imp:3s;
pas	ADV	pas	pas	01-Normé	01-Normé	_	_	_
de	PRP	de	de	01-Normé	01-Normé	_	_	_
pouvoir	NOM	pouvoir	pouvoir	01-Normé	01-Normé	m	s	_

Tableau 5 : Exemple de sortie du module d'enrichissement pour une production d'un élève de CE1

3. Quelques erreurs d'AliScol

De plus, lors du passage des fichiers dans le module d'enrichissement, j'ai repéré quelques erreurs dans les sorties de l'aligneur. Elles concernent l'étape d'étiquetage morphosyntaxique, plus précisément l'association du lemme à la forme analysée. Cela provient de *TreeTagger* et non pas d'*AliScol* car celui-ci ne corrige pas les erreurs de l'étiqueteur. Cependant, dans la tâche 2 du projet *E-Calm* il faut pouvoir travailler sur des données de la meilleure qualité possible. J'ai donc tenté de détecter les erreurs récurrentes qui pouvaient être traitées automatiquement. Certains lemmes étaient en fait la forme du mot au féminin et/ou au pluriel, par exemple « favorite » comme lemme au lieu de « favori ». Il y avait également des doubles lemmes, par exemple « maître|maîtresse » ou « folle|fou ». Il a donc fallu modifier cela pour n'avoir en lemme que la forme au masculin singulier afin d'homogénéiser les données. Ces traitements supplémentaires ont été ajoutés au module d'enrichissement.

En plus des erreurs sur le lemme, il y avait parfois des erreurs de catégorisation, probablement elles aussi dues à *TreeTagger*. J'ai donc vérifié un à un les mots étiquetés comme adjectifs et fait une liste des erreurs commises. Je n'ai cependant rien modifié à la main car ce sont des corrections à faire sur *AliScol* lui-même lors de son appel à l'étiqueteur. Il est donc plus pertinent de faire remonter les erreurs à sa conceptrice pour qu'elle puisse l'améliorer. Voici quelques exemples de mots catégorisés par erreur comme adjectifs :

« porte » et « amène » (verbes au présent), « mino » et « Coco » (noms propres), « gamine » et « tarte » (noms communs), etc.

J'ai également trouvé quelques erreurs dans le repérage des termes hypo et hyper-segmentés, par exemple « sotre » (mot attendu : « autres ») relève d'une hyper-segmentation mais n'est pas indiqué comme tel. Cela vient de l'étape d'alignement d'*AliScol* qui génère des erreurs « généralement lorsqu'une une combinaison d'erreurs apparaît, par exemple une combinaison d'erreur orthographique et d'hyper-segmentation » (Wolfarth, 2019, p. 196). Ce type de problème ne peut pas être résolu ici car les ressources en ma possession ne permettent pas de les corriger de manière automatique, il faudrait corriger l'aligneur.

Le chapitre suivant s'intéresse à la définition du modèle linguistique qui va permettre le découpage en *base* + *flexions* des données que nous venons de présenter.

Chapitre 4. Modélisation linguistique

L'enjeu de cette partie est la modélisation du découpage des formes adjectivales afin de permettre une comparaison fine entre formes produites et formes normées. Dans la littérature, la question est loin de faire l'unanimité et, selon les objectifs, on trouve plusieurs approches possibles

1. Réflexions autour de la modélisation

1.1. Enjeu de la modélisation

L'objectif de ce projet est de déterminer où ont lieu les erreurs concernant les adjectifs ; c'est-à-dire, est-ce que parmi le radical, la flexion de genre et la flexion de nombre il y a un élément qui ressort comme étant plus difficile à orthographier correctement en fonction du niveau scolaire. Il va donc falloir que je puisse positionner l'erreur à l'intérieur du mot. Pour cela je souhaite le découper en *base – flexions* afin de pouvoir pointer celle qui a posé souci à l'élève.

Une première approche simpliste pourrait être de décrire le comportement flexionnel de chaque adjectif. Cette approche permet en effet de décomposer les adjectifs normés mais elle implique une forte redondance des données. En effet, comme le montre l'exemple du Tableau 6, plusieurs adjectifs ont le même comportement flexionnel.

	Singulier	Pluriel
Masculin	joli	jolis
Féminin	jolie	jolies

	Singulier	Pluriel
Masculin	absent	absents
Féminin	absente	absentes

Tableau 6 : Exemple de deux adjectifs au même comportement flexionnel à l'écrit

L'objectif ici est donc bien de trouver une forme de modélisation permettant de proposer des solutions linguistiquement justifiées de découpage aussi bien des formes normées que non normées.

Ainsi, pour ne pas représenter deux fois la même chose, je vais chercher à créer différents modèles dont chacun correspondra à un type de comportement que partagent plusieurs adjectifs.

Maintenant, il reste à déterminer comment le faire. C'est-à-dire, comment couper les mots pour pouvoir y situer d'éventuelles erreurs ?

Avec l'exemple précédent, un premier découpage intuitif pourrait séparer la base du mot des flexions qu'on lui ajoute, comme par exemple :

joli + _ + _ *joli +_ + s* *joli + e + _* *joli + e + s*

Cependant cette approche ne propose pas de méthode pour aborder le découpage des formes erronées. Comment découper des formes telles que « heureus » ou « blache » ?

Je me suis donc renseignée sur différents modèles de découpage de mots qui existent déjà dans la littérature.

1.2. Les modèles de la littérature

1.2.1. Côté linguistique

Avant de choisir un type de modélisation, il faut déterminer ce que l'on entend par les termes base, flexion de genre et flexion de nombre. Tout d'abord, la base est la « partie du mot comprenant essentiellement la racine (base radicale) » (https://www.cnrtl.fr/definition/base), c'est-à-dire en général un lexème (Roché, 2010) soit un mot sans affixes ni marques de genre et de nombre comme par exemple « petit ».

La flexion, elle, est définie comme un « changement morphologique dans la finale d'un mot […] par l'adjonction d'un affixe ou désinence au radical » (https://www.cnrtl.fr/definition/flexion). Toutefois, dans ce travail la flexion désignera à la fois une marque et une absence de marque car pour homogénéiser les découpages il est nécessaire que chacun d'entre eux soit composé des trois parties : base, flexion de genre, flexion de nombre. L'absence de marque de genre ou de nombre sera donc indiquée par le symbole « _ ».

En effet, il n'existe ni marque de genre pour le masculin ni marque de nombre pour le singulier (Cogis, 2003 ; Laparra, 2010 ; Cogis & Brissaud, 2019). Par ailleurs, en ce qui concerne le féminin tous les linguistes ne sont pas d'accord pour lui attribuer une marque.

En effet, Cogis & Brissaud (2019) pensent que le *-e* ne devrait pas être considéré comme la marque du féminin puisqu'on le retrouve à la fin de beaucoup de verbes et de mots finissant par une consonne et qu'il permet de sonoriser cette dernière. Cependant, c'est tout

de même le choix qui a été fait ici car ce travail ne porte que sur les adjectifs. Or, en dehors de ceux qui sont épicènes et qui ne portent donc pas de marque de genre, la forme au féminin des adjectifs comporte toujours un *-e* final (Laparra, 2010).

Pour ce qui est de la flexion de nombre des adjectifs, plusieurs chercheurs s'accordent à dire que la marque écrite est toujours *-s* ou *-x* (Noailly, 1999 ; Laparra ,2010).

Maintenant que les termes base, flexion de genre et flexion de nombre ont été définis, il reste à trouver une méthode de représentation de la découpe d'un mot en ces trois parties.

1.2.2. Côté TAL

Comme présenté dans le Chapitre 2. État de l'art (cf. partie 2. Le TAL et l'analyse morphologique, p. 23), l'approche de Fradin (1994) découpe les mots en lexèmes et se compose de deux lexiques : un pour les bases et un pour toutes les formes possibles dérivées de telle ou telle base. Les lexiques sont représentés par des arbres de lettres avec les catégories syntaxiques comme nœuds (cf. Figure 10, p. 26). Ce modèle dispose également de deux types de règles : les règles morphologiques et les règles phonologiques pour analyser les mots et déterminer s'ils sont correctement construits.

Cette approche dite computationnelle à deux niveaux ne paraît pas être l'idéal dans notre cas pour plusieurs raisons. Tout d'abord, elle est composée de lexiques sous formes d'arbre de lettres ; or ce que je cherche à faire est un découpage séparant la base de la ou les flexions. Par ailleurs, les différentes catégories syntaxiques qui apparaissent dans ces arbres ne sont pas nécessaires puisque ne sont traités ici que les adjectifs. Et enfin, ce travail étant centré sur l'écrit, les règles phonologiques ne paraissent pas nécessaires (elles seraient probablement plus utiles dans le cas d'une synthèse vocale par exemple).

Le modèle de Lallich-Boidin (1987), quant à lui, découpe les mots en : *base – flexion genre – flexion nombre* selon des modèles prédéfinis. Il leur applique également des régularisations sur la base et/ou sur la forme fléchie (cf : partie 2.4.1.2 Fonction calculatoire du Chapitre 2. État de l'art, p. 28) quand le découpage ne permet pas de retrouver la bonne orthographe du mot.

Ce modèle permet mieux d'atteindre mon objectif. En effet, il permettrait bien de découper le mot en trois parties. Cependant, cette modélisation a été conçue pour être robuste

lors d'une analyse morpho-syntaxique et elle accepte donc des formes telles que « généraux » qui seront découpées en « général+_+s » et ainsi interprétées comme étant bien du masculin pluriel. Or je cherche à analyser si l'orthographe est respectée.

Par ailleurs, l'utilisation du symbole * derrière lequel se cachent plusieurs lettres différentes (un dédoublement de consonnes, l'ajout d'un « h » ou d'un « u », etc.) ne paraît pas pertinent dans ce contexte J'ai donc choisi comme alternative d'avoir des modèles comprenant parfois plusieurs bases pour un même mot afin d'éviter l'utilisation de ce symbole trop généralisant pour ce travail.

Néanmoins, cela a ajouté un problème de taille : comment définir les bases d'un mot ? Quel(s) argument(s) prendre en compte pour découper le mot de telle ou telle manière ? Ces questions sont très importantes car ce sont leurs réponses qui détermineront quelle interprétation faire de tel ou tel cas.

1.3. Approche choisie

Lors des diverses réunions de la tâche 2 avec M. Ponton, Mme Brissaud et d'autres membres du projet E-Calm, nous avons longuement réfléchi à la façon de modéliser car plusieurs cas étaient sujets à controverse. En effet, nous étions tous d'accord pour des découpages tels que celui du mot « absent » qui étaient assez intuitifs (cf. Tableau 7).

Genre	Nombre	Base	Flexion genre	Flexion nombre
Masculin	**Singulier**	absent	_	_
	Pluriel	absent	_	s
Féminin	**Singulier**	absent	e	_
	Pluriel	absent	e	s

Tableau 7 : Description du découpage pour le mot « absent »

Mais les cas possédant un dédoublement de consonne lorsque l'on passe du masculin au féminin ne faisaient pas l'unanimité. La question était : la deuxième consonne doit-elle être intégrée à la base du mot (solution 1) ou bien à la flexion de genre (solution 2) ?

Si nous attendons la forme « gentil », les différents découpages et interprétations en fonction de la forme produite et de la solution choisie sont ceux présentés dans le Tableau 8.

		Découpage		**Interprétation**	
Cas	**Mot produit**	**Solution 1**	**Solution 2**	**Solution 1**	**Solution 2**
A	gentil	gentil + _ + _		Orthographe correcte	
B	gentile	gentil + e + _	genti + le + _	Erreur genre	Erreur base + genre
C	gentill	gentill + _ + _		Erreur base	
D	gentille	gentill + e + _	gentil + le + _	Erreur base + genre	Erreur genre

Tableau 8 : Possibilités de découpages et interprétations de différentes formes produites du mot « gentil »

Si nous attendons la forme « bonne », les différents découpages et interprétations en fonction de la forme produite et de la solution choisie sont ceux présentés dans le Tableau 9.

		Découpage		**Interprétation**	
Cas	**Mot produit**	**Solution 1**	**Solution 2**	**Solution 1**	**Solution 2**
E	bonne	bonn + e + _	bon + ne + _	Orthographe correcte	
F	bonn	bonn + _ + _	bonn +_ + _	Erreur de genre	Erreur base + genre
G	bon	bon + _ + _		Erreur base + genre	Erreur genre
H	bone	bon + e + _	bo + ne + _	Erreur base	Erreur base

Tableau 9 : Possibilités de découpages et interprétations de différentes formes produites du mot « bonne »

Il faut savoir que ces découpages partent de la fin du mot pour aller à son début, comme expliqué dans la Figure 11 (p. 28) ; c'est-à-dire que c'est d'abord la flexion de nombre qui est extraite, puis celle de genre pour ensuite obtenir la base. Ceci est important à garder en tête afin de comprendre, par exemple, le découpage du cas B par la solution 2 qui est « genti + le + _ » et non pas « gentil + e + _ ».

Le choix d'un découpage de droite à gauche a été fait car l'inverse ne permet pas toujours de pouvoir examiner les flexions et la base séparément. En effet, si la base du mot est erronée il devient difficile de trouver une interprétation à la flexion obtenue. Si par exemple l'élève produit le mot « gentill » au lieu de « gentil », un découpage de gauche à droite donnerait « gentil + l » et déduire quelque chose de la flexion « l » ou même simplement pouvoir la catégoriser comme marque de genre ou de nombre devient compliqué.

Ainsi, les deux solutions impliquent des interprétations différentes qui ont chacune leurs avantages et leurs inconvénients. En effet dans les cas D et G avec la solution 1, il peut être embêtant de considérer qu'il y a à la fois une erreur de genre et de nombre alors que le mot orthographié tel quel existe et qu'il ne respecte simplement pas le genre attendu : ici la

solution 2 semble plus adaptée. À l'inverse, dans le cas B avec la solution 2, déduire une erreur de base et de genre alors que la différence entre la forme attendue et celle produite est l'ajout d'un *-e* final peut paraître excessif : ici la solution 1 semble plus appropriée.

Afin de pouvoir choisir entre les deux solutions l'équipe et moi-même nous sommes recentrés sur le sujet : nous étudions l'orthographe des mots et nous concentrons donc sur l'écrit et non sur l'oral. C'est pour cela que nous avons décidé que le *-e* final resterait la marque du féminin dans ce travail et nous avons donc appliqué la solution 1 lors de la modélisation.

Concernant les mots tels que « égal », nous avons fait le choix de désigner *-ux* comme flexion de nombre et non pas seulement *-x* comme dans le cas de « beau ». En effet, historiquement, on a utilisé le x comme équivalent de *-us* puis cela été oublié et un « u » a été ajouté devant le *-x*. Le « u » fait donc également partie de ce qui est ajouté à la base du mot dans tous les cas pluriels de ce type, exemple : « général – généraux », « global – globaux », etc.) alors qu'il est compris dans la forme au singulier dans le cas de « beau ».

Par ailleurs, les mots comme « bel », « fol », « nouvel » et « vieil » n'ont pas été intégrés dans les modèles et seront traités comme des cas particuliers par l'outil de découpage que je vais présenter plus loin (cf. Chapitre 5. Conception *d'AliAdj*, p. 53). En effet, ce sont des formes avec un comportement unique : la modélisation ne leur apporterait donc rien.

Une fois ces questions résolues, l'ensemble des modèles de découpage flexionnel des formes adjectivales peut être défini et ainsi permettre d'envisager le découpage des formes non normées.

2. Construction du modèle

Le modèle construit dans les parties suivantes respecte donc les choix effectués précédemment, cela implique que :

- Chaque mot doit être découpé en trois parties : base – flexion de genre – flexion de nombre.

- La flexion de nombre peut être *-s*, *-x* ou *-ux* en fonction du cas sauf pour les adjectifs invariables où elle est représentée par « _ ».
- La flexion du genre est *-e* pour le féminin et « _ » pour le masculin et les adjectifs épicènes.
- La base est ce qu'il reste lorsque les flexions de genre et de nombre ont été séparées du mot.
- Il peut être nécessaire d'avoir plusieurs bases pour un même lemme. Par exemple, le lemme « beau » a comme base au masculin « beau » et comme base au féminin « bell ».
- Les découpages s'effectuent de la droite vers la gauche.

Avant de pouvoir modéliser tout ceci, il a d'abord été nécessaire d'effectuer ces découpages sur les formes produites afin de pouvoir se rendre compte des regroupements possibles des adjectifs qui présentent le même comportement flexionnel.

2.1. Découpage des formes produites

Ainsi, une fois les données récupérées et enrichies comme décrit dans le chapitre précédent (cf. Chapitre 3. Prétraitement des données, p. 35), j'ai répertorié tous les adjectifs des différentes productions afin d'avoir toutes les formes possibles écrites par les élèves. Je les ai ensuite listés comme indiqué dans le Tableau 10.

Forme attendue	Nombre d'occurrences	Nombre d'erreurs	Formes produites erronées	Nombre d'apparition par forme
autre	31	8	lotre	3
			autres	2
			nautre	1
			aute	1
			otre	2
autres	17	11	autre	7
			dotre	3
			dautre	1

Tableau 10 : Exemple de la liste des formes produites pour « autre » et « autres » en CE2

Cette étape m'a permis d'observer les erreurs fréquemment commises par les élèves. Pour la suite, j'ai sélectionné dans ces listes uniquement les formes produites n'étant pas catégorisées par l'aligneur comme : hypo-segmentées, hyper-segmentées, non-

pertinentes ou omises. Cette sélection était nécessaire car le découpage de ces formes aurait été soit impossible soit bien plus complexe. En effet, il aurait fallu faire un découpage supplémentaire pour récupérer le mot lors d'une hypo-segmentation (ex : « touteseulle ») ainsi qu'une recherche et un assemblage lors d'une hyper-segmentation (ex : « cents tième »). De plus, les mots classés comme non-pertinents sont souvent bien loin de la forme attendue (ex : « s' » au lieu de « seul », « ere » à la place de « fiers » ou encore « sardre» à la place de « sourd ») ; or cela ne permet pas de représenter la capacité de l'élève à orthographier le mot voulu.

J'ai également retiré les formes dont le lemme était « <unknown> » car celles-là ne possèdent pas d'information de genre et de nombre. Or, ces traits morphologiques sont indispensables à connaitre pour pouvoir effectuer un découpage automatique des mots. Le Tableau 11 montre la proportion des adjectifs qui sont donc non traitables ici.

	CP	CE1	CE2	CM1	3°	Licence1	Total	Total (%)
Hypo-segmentés	20	29	20	23	0	0	**92**	**1,70%**
Hyper-segmentés	12	14	9	15	0	0	**50**	**0,92%**
Non-pertinents	5	11	8	5	0	0	**29**	**0,54%**
Omis	1	2	3	14	4	0	**24**	**0,44%**
Lemme inconnu	2	23	52	83	63	10	**233**	**4,31%**
Total	**40**	**79**	**92**	**140**	**67**	**10**	**428**	**7,91%**

Tableau 11 : Quantité d'adjectifs exclus du traitement en fonction du type de problème.

On remarque dans le tableau ci-dessus que le taux d'adjectifs retirés correspond à 7,91% du total des adjectifs. En ramenant ces chiffres à la globalité du corpus (cf : Tableau 2, p. 15), les résultats du Tableau 12 montrent qu'il s'agit de 0,29% des données.

	CP	CE1	CE2	CM1	3°	Licence1	Total	Total (%)
Adjectifs acceptés	374	798	1394	1753	525	135	**4979**	**3,36%**
Adjectifs refusés	40	79	92	140	67	10	**428**	**0,29%**
Total	**414**	**877**	**1486**	**1893**	**592**	**145**	**5407**	**3,65%**

Tableau 12 : Nombre d'adjectifs par rapport au nombre total de tokens (calculés dans le Tableau 16)

Les adjectifs qui pourront donc être traités dans ce travail représentent 3,36% de l'ensemble du corpus disponible.

Ainsi, pour passer du Tableau 10 au Tableau 13, j'ai retiré les formes « lotre » et « dautre ». Une fois cette sélection effectuée je me suis appuyée sur la modélisation choisie et décrite précédemment pour découper la forme produite en trois parties : *base + flexion de*

genre + flexion de nombre. Pour cela, j'ai séparé le mot en allant de la droite vers la gauche : il faut d'abord extraire la marque de nombre si elle est présente, ensuite celle de genre et enfin il reste la base. En suivant ces règles précises on obtient le découpage présenté dans le Tableau 13.

Forme attendue	Forme produite	Découpage produit	Erreur radical	Erreur flexion genre	Erreur flexion nombre
autre	autres	autre + _ + s	non	non	oui
	nautre	nautre + _ + _	oui	non	non
	aute	aute + _ + _	oui	non	non
	otre	otre + _ + _	oui	non	non
autres	autre	autre + _ + _	non	non	oui
	dotre	dotre + _ + _	oui	non	oui

Tableau 13 : Exemple de la liste des découpages des formes produites pour « autre » et « autres » en CE2

Il reste des formes hypo-segmentées dans le tableau ci-dessus (« nautre » et « dotre ») car elles n'ont pas été indiquées comme tel par *AliScol*.

2.2. Extraction des modèles de comportement flexionnel

À la suite de ces découpages, j'ai classé les adjectifs en fonction de leurs variations morphologiques afin de pouvoir résumer leurs comportements en une liste de modèles. Le Tableau 14 est un extrait de ces regroupements.

Comportement 1	Comportement 2	Comportement 3	Comportement 4	…
absent	adulte	additionnel	affreux	…
adolescent	agile	émotionnel	amoureux	
âgé	agricole	éternel	avantageux	
aîné	angora	gentil	baveux	
aisé	arrière	immortel	boueux	
…	…	…	…	…

Tableau 14 : Extrait de la liste des adjectifs regroupés selon leur comportement flexionnel

Ici, le comportement représente les variations les plus fréquentes, c'est-à-dire ajout du *-s* pour le pluriel, du *-e* pour le féminin et base identique au lemme. Le comportement 2 regroupe les adjectifs épicènes, donc sans marque de genre. Le comportement 3, quant à lui, rassemble les mots avec dédoublement de consonne lors du passage au féminin ; il représente donc des adjectifs avec deux bases différentes (une pour le masculin et une pour le féminin). Il en va de même pour le comportement 4 qui décrit des adjectifs à bases différentes en

fonction du genre. Cependant, ici le lemme perd son *-x* et gagne un *-s* en passant au féminin (« amoureux + _ + _ » vs. « amoureus + e + _ »). La liste des comportements flexionnels ne s'arrête pas là, mais ils seront tous représentés dans la modélisation qui suit.

De ces comportements, j'ai pu extraire 29 modèles, validés par les membres du projet E-Calm. Cependant, après réflexion nous avons choisi d'ajouter le modèle « vengeur » car même si aucun adjectif relevé ici ne lui correspond, c'est une forme de variation que l'on peut trouver en français. Or, si nous souhaitons que l'outil de découpage puisse être utilisé sur d'autres corpus, il doit couvrir tous les comportements flexionnels adjectivaux possibles. J'ai donc finalement un total de 29 modèles (cf. Annexe 6, p. 95), en voici quelques exemples dans le Tableau 15.

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	égal + 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 15 : Extrait de la liste des modèles et des découpages correspondants

Si l'on se réfère aux comportements flexionnels décrits précédemment (Tableau 14) : dans cette liste « absent » modélise le comportement 1, « agile » le comportement 2, « additionnel » le comportement 3 et « affreux » le comportement 4.

On remarque une réelle ressemblance entre ces modèles et la modélisation du CRISS sur laquelle ce travail s'est appuyé. Cependant, la différence ici est l'utilisation de bases multiples. En effet, nous ne souhaitions pas que des formes telles que « égal + _ + s » puissent être acceptées. Dans cette modélisation, « égal » possède deux bases : « éga » pour le masculin pluriel et « égal » pour le reste (cf. modèle « égal », Tableau 15). De ce fait, le découpage attendu du masculin pluriel est « éga + _ + ux » et celui de la forme erronée est : « égals + _ + _ ». Tandis que la forme erronée est découpée en « égal + _ + s » par le modèle

du CRISS qui accepte bien cela comme un masculin pluriel. A l'inverse, notre modélisation interprète « égals » comme une erreur de base et de flexion de nombre ; ce qui est exactement ce que l'on souhaite ici.

Par ailleurs, la forme non normée « heureus » dont il était question dans la partie 1.1 Enjeu de la modélisation (p. 41) pourra être découpée et interprétée différemment en fonction de l'attendu :

- Si le masculin singulier ou le masculin pluriel est attendu, « heureus + _ + _ » implique une erreur de base.
- Si le féminin singulier est attendu, « heureu + _ + s » implique une erreur de base, de flexion de genre et de flexion de nombre.
- Si le féminin pluriel est attendu, « heureu + _ + s » implique une erreur de base et de flexion de genre.

De plus la forme « gentill » à la place de « gentille » qui illustrait précédemment (p.45) l'importance du découpage de gauche à droite est découpée ici (de droite à gauche) en « gentill + _ + _ » et indique une erreur de base.

Les formes erronées ne sont donc pas si évidentes à découper, c'est pourquoi il est primordial d'avoir une modélisation précise afin de permettre au module de traitement des adjectifs d'effectuer un découpage cohérent même lorsqu'il rencontre une forme inconnue.

Partie 3

-

AliAdj : module de traitement des formes adjectivales

Chapitre 5. Conception *d'AliAdj*

La modélisation du comportement des adjectifs terminée, je m'intéresse maintenant à la conception du module de traitement de l'adjectif nommé *AliAdj*. Ce système a pour but de découper des formes normées et des formes produites selon le modèle linguistique présenté dans le chapitre précédent (cf. Tableau 15, p. 50) et de les comparer afin de trouver où est localisée l'erreur s'il y en a. Pour cela, il reste encore à définir les corpus sur lesquels travailler et à adapter les modèles et le lexique conçus plus tôt aux traitements informatiques.

1. Définition des Corpus

1.1. Conception des corpus

Afin de pouvoir construire et évaluer *AliAdj*, il a d'abord été nécessaire d'extraire deux échantillons de l'ensemble des données. L'un appelé corpus de référence servira à l'évaluation du système et l'autre appelé corpus de travail servira à construire cet outil.

J'ai donc récupéré 10% des productions par niveau scolaire de façon aléatoire pour construire le corpus de référence et j'ai réitéré cette étape pour construire le corpus de travail ; ainsi les deux échantillons sont différents mais approximativement de la même taille. Le Tableau 16 présente le nombre de productions par niveau et par corpus ainsi que le nombre de tokens utilisables[39].

	Corpus général		Corpus de travail		Corpus de référence	
	Productions	**Tokens**	**Productions**	**Tokens**	**Productions**	**Tokens**
CP	337	9 632	34	1 024	34	1 005
CE1	337	24 226	34	2 672	34	2 604
CE2	337	43 539	34	4 349	34	4 192
CM1	337	53 440	34	5 548	34	5 381
3ème	43	14 450	4	1 086	4	933
Licence1	23	2 864	2	228	2	321
Total	1414	148 151	142	14 907	142	14 436

Tableau 16 : Totaux des productions et tokens en fonction du type de corpus et du niveau scolaire

[39] C'est-à-dire les tokens qui ont pu être catégorisés par l'aligneur présenté dans la partie 1. Alignement des productions du Chapitre 3. Prétraitement des données, p. 35.

Les chiffres entre les deux sous-corpus ont l'air similaires. Cependant, pour vérifier la significativité de ces échantillons, j'ai calculé le pourcentage d'adjectifs présents dans chaque corpus par rapport au nombre de tokens. Les résultats sont présentés dans le Tableau 17.

	CP	CE1	CE2	CM1	3°	Licence 1	Total
Corpus général	4,3%	3,62%	3,41%	3,54%	4,1%	5,06%	3,65%
Corpus de travail	4,0%	2,99%	3,04%	3,51%	2,21%	2,63%	3,21%
Corpus de référence	4,28%	4,07%	3,27%	3,2%	3,97%	5,61%	3,55%

Tableau 17 : Taux d'adjectifs trouvés dans les productions en fonction du corpus et du niveau scolaire

La différence entre le pourcentage d'adjectifs présents au niveau de 1ère année de licence dans le corpus de travail et dans les autres corpus peut s'expliquer par le fait que cet échantillon est plus petit que ceux des autres niveaux : il ne possède pas autant de productions d'élèves (cf : Tableau 16). La variabilité inter-élèves peut donc être plus visible qu'avec un échantillon plus grand.

On remarque cependant, que même en tenant compte de cet écart, la différence de taux d'adjectif du corpus général par rapport aux sous-corpus n'excède pas 2,5%. De plus, les pourcentages d'adjectifs totaux de chaque corpus sont très proches (moins de 0,3% d'écart). On peut donc dire que les échantillons sont bien représentatifs du corpus général.

1.2. Annotation du corpus de référence

Maintenant que la représentativité des sous-corpus a été démontrée, il faut annoter le corpus de référence. C'est en effet lui qui va servir d'outil de comparaison pour évaluer le module de traitement des adjectifs. Pour obtenir une qualité optimale des données, j'ai annoté le corpus de référence manuellement. Pour une qualité encore meilleure ainsi que pour évaluer le modèle de découpage, il aurait fallu que cette annotation soit réalisée par plusieurs annotateurs afin de comparer les résultats inter-annotateurs. Pour des raisons de temps, ce travail n'a pu être mené à bien et pourra être mené dans une deuxième phase d'évaluation.

Pour réaliser cette annotation, j'ai d'abord extrait les adjectifs des données, je les ai ensuite découpés en suivant le modèle linguistique élaboré dans la partie précédente (cf. Annexe 6, p. 95). Je n'ai fait ce travail que pour les adjectifs qui ont été correctement catégorisés en tant que tel. En effet, on trouve parfois des mots indiqués comme adjectifs mais qui sont en fait des noms si l'on tient compte de leur contexte d'apparition.

Ce découpage est représenté à travers six colonnes ajoutées aux fichiers : « BaseAtt », « FlexGenreAtt », « FlexNbAtt », « BaseProd », « FlexGenreProd » et « FlexNbProd ». Les trois premières montrent le découpage du mot attendu et les trois suivantes montrent le découpage du mot produit. J'ai également ajouté les trois colonnes suivantes : « ErrBase », « ErrFlexGenre », et « ErrFlexNb ». Elles indiquent les trois types d'erreurs possibles : sur la base, sur la flexion de genre et sur la flexion de nombre. Le symbole « _ » qui peut être présent dans les colonnes indiquant les flexions de genre et de nombre montre une absence de marque. Tandis que les trois dernières colonnes, elles, ne peuvent avoir comme valeur que le chiffre « 1 » pour une erreur et le chiffre « 0 » pour une absence d'erreur.

Le Tableau 18 est un exemple de cette annotation et représente également ce à quoi doivent ressembler les fichiers produits par l'outil *AliAdj*. Dans un souci de présentation, j'ai supprimé dans cet exemple toutes les colonnes ne présentant pas d'intérêt dans l'explicitation de cette étape.

Seg Norm	Seg Trans	Genre	Nombre	BaseAtt	Flex Genre Att	Flex Nb Att	Base Prod	Flex Genre Prod	Flex Nb Prod	Err Base	Err Flex Genre	Err Flex Nb
heureux	eurheux	m	_	heureux	_	_	eurheux	_	_	1	0	0
petit	petit	m	s	petit	_	_	petit	_	_	0	0	0
gentille	gentille	f	s	gentill	e	_	gentill	e	_	0	0	0
noires	noir	f	p	noir	e	s	noir	_	_	0	1	1
meilleurs	meilleur	m	p	meilleur	_	s	meilleur	_	_	0	0	1

Tableau 18 : Exemple de l'annotation du corpus de référence, ici production d'un élève de CE1 (ajout des neuf dernières colonnes)

2. Modélisation informatique

2.1. Algorithme d'AliAdj

L'outil *AliAdj* doit pouvoir produire un fichier csv ayant le même format que celui du corpus de référence annoté (cf. Tableau 18). Il doit donc découper et comparer toutes les formes adjectivales produites.

Cela nécessite de posséder certaines données en entrée du système qui sont les suivantes : les fichiers des productions (un par niveau scolaire) sortis du système *AliScol* et complétés par le module d'enrichissement, un fichier présentant les modèles des

comportements flexionnels des adjectifs et un fichier « lexique » associant chaque lemme adjectival à un modèle de comportement morphologique.

Une fois toutes ces ressources récupérées le système va devoir parcourir chaque ligne de chaque fichier de production (une ligne correspond à un adjectif) et suivre l'algorithme suivant (ces étapes seront illustrées avec un exemple dans la partie 2.3 Déroulement du système, p. 59) :

- Récupérer le lemme, les formes attendue et produite, le genre, le nombre et le statut de segmentation.
- Vérifier que l'adjectif est traitable : son statut de segmentation ne doit pas être hypo-segmenté, hyper-segmenté, non pertinent ou omis et son lemme ne doit pas être marqué « <unkown> ».
- Chercher dans le lexique à quel modèle de comportement correspond le lemme de la forme produite.
- Chercher dans les modèles quel est le comportement attendu en fonction du lemme de la forme produite, de son genre et de son nombre.
- Découper les formes normée et produite en fonction du modèle sélectionné précédemment.
    - Chercher la marque de nombre possible (selon le modèle) sur la forme produite. Si elle y est et qu'elle doit y être alors le nombre est marqué comme normé mais si elle ne devrait pas y être alors le nombre est marqué comme erroné. Dans les deux cas, la marque du nombre est alors mise de côté sur les formes produite et normée afin qu'il ne reste que les bases et les flexions de genre.

      S'il n'y a aucune marque de nombre, le mot à analyser n'est pas modifié et la flexion de nombre est marquée comme normée ou erronée en fonction de si elle était attendue ou non.
    - Réitérer cette étape pour le genre : chercher la marque de genre. Si elle est présente et qu'elle doit être là alors le genre est normé, à l'inverse si elle devait être absente alors le genre est erroné. Dans les deux cas, la marque de genre est séparée des formes produite et normée afin qu'il ne reste que les bases.

      S'il n'y a aucune marque de genre, le mot à analyser n'est pas modifié et la flexion de genre est marquée comme normée ou erronée en fonction de si elle était attendue ou non.

    - Comparer la base de la forme normée à celle de la forme produite : si elles sont identiques la base est normée sinon elle est erronée.
- Ecrire le résultat de l'analyse dans l'ordre suivant : base, flexion de genre et flexion de nombre de la forme normée, base, flexion de genre et flexion de nombre de la forme produite, erreur sur la base, erreur sur la flexion de genre, erreur sur la flexion de nombre. S'il y a erreur elle est marquée « 1 » sinon « 0 ».
- Passer à l'adjectif suivant ou au fichier suivant.

Tout ce traitement requiert des données spécifiques soit les fichiers de productions, une liste des modèles de comportements morphologiques de l'adjectifs et un lexique pour associer ces modèles aux adjectifs.

2.2. Lexique et modèles de comportement

Les fichiers de productions ayant déjà été mis en forme grâce à *AliScol* et complétés par le module d'enrichissement, il reste à créer le lexique et les modèles de comportement.

Au fur et à mesure de ce travail, j'ai donc conçu un lexique spécifique[40] aux adjectifs. C'est un fichier « .txt » qu'il faut modifier lorsque l'on souhaite traiter de nouvelles données : le lemme de chaque nouvel adjectif ainsi que son modèle de variation morphologique doivent y être indiqués de la façon présentée dans le Tableau 19.

lemme	**modèle**
absent	absent
accompli	absent
acéré	absent
additionnel	ancien
adolescent	absent
adorable	agile
adulte	agile
affamé	absent
affreux	affreux
âgé	absent

Tableau 19 : Extrait du lexique associant chaque lemme à un modèle de comportement morphologique

[40] Disponible à l'adresse suivante : https://github.com/RachelGaubil/Outiller_description_morpho_adjectivale_primaire_universite/tree/master/AliAdj/modelisation

En ce qui concerne les modèles de comportement, j'en ai diminué le nombre en réadaptant la modélisation linguistique faite précédemment (cf. partie 2.2 Extraction des modèles de comportement flexionnel du Chapitre 4. Modélisation linguistique, p. 49) : j'ai regroupé les cas similaires. Le nombre de modèles est passé de 29 à 27.

Ainsi, j'ai regroupé les modèles « additionnel » et « muet » sous « ancien » qui présentent le même comportement flexionnel à la différence près que la consonne à doubler dans la forme au féminin n'est pas la même (« l », « t » ou « n ») ; c'est ce qui explique leur indépendance dans la modélisation linguistique. Cependant, dans le modèle informatique cela ne pose aucun problème de les regrouper sous une catégorie présentant un dédoublement de la consonne finale lorsque le mot est mis au féminin puisqu'ils seront traités de la même façon.

Au-delà du nombre de modèles, cette nouvelle modélisation[41] diffère de la précédent (cf. Annexe 6, p. 95) au niveau du format : on a ici un fichier « .txt » contenant des informations plus lisibles pour le système. La différence est qu'elle présente une ligne pour chaque découpage et chaque modèle est donc représenté par plusieurs lignes. Le Tableau 20 est un extrait du fichier.

Lemme	Base	FlexGenre	FlexNb	GenreNb
absent	absent	_	_	m,s
absent	absent	_	s	m,p
absent	absent	e	_	f,s
absent	absent	e	s	f,p
bas	bas	_	_	m,_
bas	bass	e	_	f,s
bas	bass	e	s	f,p
bref	bref	_	_	m,s
bref	bref	_	_	m,p
bref	brèv	e	_	f,s
bref	brèv	e	s	f,p

Tableau 20 : Exemple de la liste des modèles adaptée pour *AliAdj*

Les colonnes « FlexGenre » et « FlexNb » indiquent la marque de genre et/ou de nombre qu'il faut ajouter à la base du mot pour obtenir la forme voulue tandis que le symbole « _ » représente l'absence de cette marque. La colonne « GenreNb » donne le genre et le

[41] Disponible à l'adresse suivante : https://github.com/RachelGaubil/Outiller_description_morpho_adjectivale_primaire_universite/tree/master/AliAdj/modelisation

nombre du découpage actuel. Attention, le symbole « _ » ici change de sens : il indique une invariabilité en genre et/ou en nombre :

- « _,s » ou « _,p » : invariabilité en genre,
- « m,_ » ou « f,_ » : invariabilité en nombre,
- « _,_ » : invariabilité totale.

La composition du lexique et des modèles de comportement ayant été vus plus en détail, je vais illustrer l'algorithme décrit précédemment à l'aide d'un exemple.

2.3. Déroulement du système

Voici donc la description du fonctionnement du module de traitement des adjectifs[42]. Pour sa conception et son entraînement, je me suis servie du corpus de travail décrit précédemment, du lexique construit au fur et à mesure du projet ainsi que de la liste des modèles réduite présentée ci-dessus. Pour éclaircir les explications qui vont suivre, elles seront illustrées au fur et à mesure par le même exemple : « âgée » en tant que forme attendue et « agées » comme forme produite.

Avant toute chose, ne sont sélectionnés dans les données que les adjectifs afin de réduire la quantité d'informations fournie au module.

Ce programme prend donc en entrée le lexique qui associe chaque lemme à un modèle et le fichier de modélisation qui associe à chaque modèle un comportement morphologique. L'entrée principale du système reste cependant les données à traiter : soit un ensemble de productions d'élèves sous format « .csv » qui ont déjà été passées dans le système *AliScol* et complétées par le module d'enrichissement (ajout du genre et du nombre pour les adjectifs), la Figure 17 représente le schéma global du fonctionnement *d'AliAdj* qui sera explicité au fur et à mesure.

[42] Disponible à l'adresse suivante : https://github.com/RachelGaubil/Outiller_description_morpho_adjectivale_primaire_universite/tree/master/AliAdj

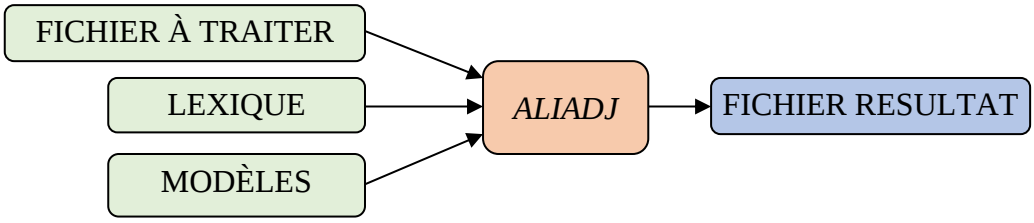

Figure 17 : Schéma des entrées et sortie *d'AliAdj*

Une fois lancé, le système récupère les informations suivantes : lemme, forme attendue, forme produite, genre, nombre et le statut de segmentation. *AliAdj* commence par vérifier le statut de segmentation du mot (cf. Figure 18). En effet, comme expliqué plus tôt (cf. partie 2.1 Découpage des formes produites du Chapitre 4. Modélisation linguistique, p. 47), les mots catégorisés comme hyper-segmentés, hypo-segmentés, non pertinents et omis ne seront pas traités. Il en est de même pour les formes dont le lemme est « <unknown> ».

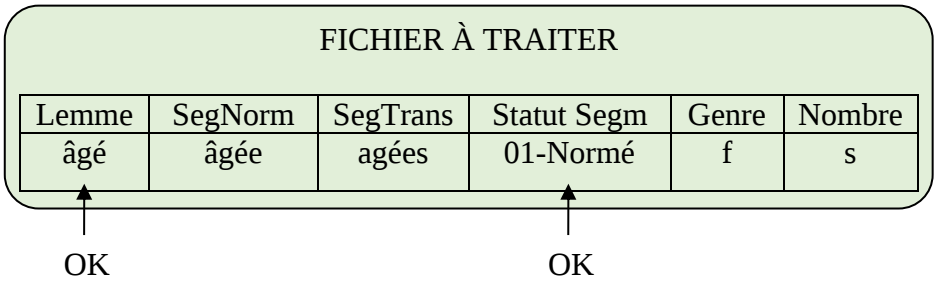

Figure 18 : Schéma des deux premières étapes *d'AliAdj*

Si ces conditions sont respectées alors le système cherche dans le lexique à quel modèle est associé le lemme de la forme étudiée. Lorsque c'est fait, il cherche dans le modèle correspondant quel est le comportement attendu grâce aux informations de genre et de nombre (cf. Figure 19).

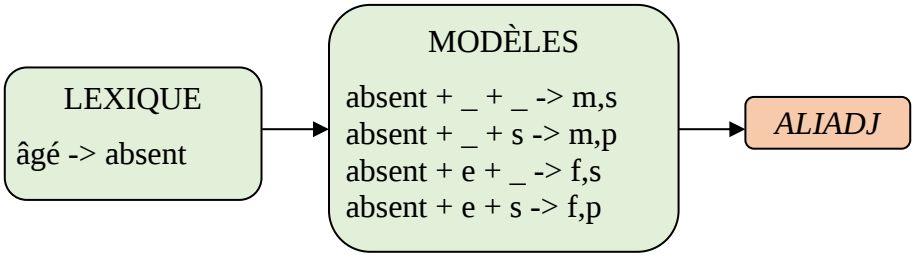

Figure 19 : Exemple des données nécessaires récupérées dans le lexique et les modèles

Une fois cela défini, il va découper la forme produite au fur et à mesure afin de pouvoir effectuer une comparaison morceau par morceau entre celle-ci et le découpage attendu. Ce découpage se fait de la droite vers la gauche : flexion de nombre, flexion de genre et enfin base.

Pour cela, le système regarde si la forme attendue possède une flexion de nombre : si oui, il va noter sa présence ou non dans la forme produite. Si aucune marque de nombre n'est attendue : il va déterminer si la forme produite contient tout de même la flexion de nombre du mot ; si c'est le cas il va signaler une erreur de nombre. S'il y en avait une, il écarte la marque de nombre pour réitérer ces étapes avec la flexion de genre. Il écarte ensuite la flexion de genre si elle était présente sur la forme produite afin qu'il ne lui reste plus que la base du mot. Une fois que c'est le cas il compare donc cette base à celle de la forme normée afin de déterminer s'il y a une erreur sur la base du mot ou non. Toutes ces étapes sont illustrées par la Figure 20.

ALIADJ

Comparaisons	**Forme attendue**	**Forme produite**	**Interprétation**
1 : flexion nombre	âgée + _	agée + s	Erreur nombre
2 : flexion genre	âgé + e	agé + e	Genre normé
3 : base	âgé	agé	Erreur base

Figure 20 : Comparaisons effectuées par *AliAdj* dans l'ordre chronologique

Le fichier produit en sortie de ce module respecte le même format[43] que le fichier décrit lors de la conception du corpus de référence (cf. Tableau 18, p. 55). C'est-à-dire qu'il conserve le format et les données du fichier à traiter et ajoute à la suite les colonnes donnant le découpage de la forme attendue, celui de la forme produite ainsi que le type d'erreurs commises s'il y en a : « 1 » pour une erreur et « 0 » pour une absence d'erreur (cf. Figure 21).

FICHIER RESULTAT

Lemme	SegNorm	SegTrans	Statut Segm	Genre	Nombre	BaseAtt	FlexGenreAtt
âgé	âgée	agées	01-Normé	f	s	âgé	e

FlexNbAtt	BaseProd	FlexGenreProd	FlexNbProd	ErrBase	ErrFlexGenre	ErrFlexNb
_	agé	e	s	1	0	1

Figure 21 : Exemple de sortie *d'AliAdj*

[43] La Figure 21 ne respecte pas tout à fait ce format, car pour que toutes les informations importantes ici puissent apparaître certaines colonnes ont été supprimées.

Il y a cependant quelques exceptions. En effet, les mots « bel », « fol », « nouvel » et « vieil » n'ont pas été modélisés à cause de leur comportement unique. C'est pourquoi ils sont traités par *AliAdj* comme des cas particuliers et ne sont donc pas recherchés dans le lexique. Hormis cela, le fonctionnement du système reste identique pour tous les autres cas. Par ailleurs, dans le cas où le lemme n'est pas trouvé dans le lexique, *AliAdj* ne produit aucun découpage. Cela signifie soit que le mot n'est pas un adjectif soit qu'il faut ajouter ce lemme et son modèle correspondant dans le lexique.

Chapitre 6. Evaluation d'*AliAdj*

1. Technique d'évaluation

Évaluer l'outil conçu ici est une étape importante du travail de recherche car c'est en démontrant la qualité du système qu'il devient possible de prouver que les objectifs ont été atteints (Popescu-Belis, 2007). En effet, cela permet d'avoir une idée précise des capacités de traitement et ainsi de ne pas sous-évaluer ou surévaluer les résultats.

L'étape d'évaluation fait également ressortir la localisation des erreurs de traitement et permet ainsi de pouvoir améliorer l'outil.

Au fur et à mesure de la conception du module, je l'ai testé sur le corpus de travail afin de pouvoir corriger les éventuels bugs, erreurs et oublis dont il pourrait faire preuve. Une fois qu'il m'a paru satisfaisant, il a été lancé sur le corpus de référence afin d'être évalué. En effet, habituellement les systèmes d'alignements sont évalués en comparant les sorties du système avec des données de références construites manuellement (Wolfarth, 2019) et qui sont différentes des données à l'aide desquelles l'outil a été construit. Le résultat du système a été comparé à l'annotation manuelle effectuée précédemment sur le corpus de référence (cf. partie 1.2 Annotation du corpus de référence, p. 54).

La métrique F-mesure semblent être particulièrement bien indiquée dans cette évaluation car elle permet d'identifier les éléments pertinents parmi l'ensemble des éléments (Popescu-Belis, 2007). De plus, les mesures de précision et de rappel que combine la F-mesure font aujourd'hui l'objet d'un certain consensus entre les chercheurs (Kraif, 2001, cité par, Wolfarth, 2019). J'ai donc choisi de déterminer les taux de rappel (R) et de précision (P) de l'outil ainsi que la F-mesure (F) *d'AliAdj* dont les formules sont les suivantes :

$$R = \frac{VP}{VP + FN} \qquad P = \frac{VP}{VP + FP} \qquad F = 2 \times \frac{P \times R}{P + R}$$

Le rappel représente l'exhaustivité du système en donnant le rapport entre le nombre de fois où le mot est découpé par rapport au nombre de fois où il aurait dû l'être. La précision, elle, indique le rapport entre le nombre de bons découpages (découpage attendu et correct) et le nombre total de découpages ; elle évalue l'exactitude du système. La F-mesure est la moyenne harmonique des deux indicateurs précédents, c'est elle qui détermine la qualité, la pertinence du système.

Pour calculer ces indicateurs, j'ai comparé ligne par ligne le corpus de référence annoté à la main avec les résultats de ce même corpus passé dans le système. Le Tableau 21 présente le comptage des adjectifs en fonction du cas dans lequel ils se trouvent.

Cas	Signification	Nombre d'occurrences
VP : vrai positif	mot à découper + est bien découpé	472
VN : vrai négatif	mot à ne pas découper + n'est pas découpé	36
FP : faux positif	mot à ne pas découper + est découpé	2
FN : faux négatif	mot à découper + n'est pas découpé	3

Tableau 21 : Nombre d'occurrences des découpages (réussis ou non) en fonction de ce qui était attendu

J'ai ensuite appliqué les formules précédentes avec les résultats obtenus dans ce tableau.

$$R = \frac{472}{472 + 3} = 0{,}9937 \qquad P = \frac{472}{472 + 2} = 0{,}9958 \qquad F = 2 \times \frac{0{,}9958 \times 0{,}9937}{0{,}9958 + 0{,}9937} = 0{,}9947$$

Avec ces chiffres on observe donc des taux de rappel et de précision tous deux à environ 99% et une F-mesure de 99% également. Avec ces pourcentages, le système paraît très performant et pertinent. Cependant les chiffres sont étonnamment élevés.

Le 1% restant comme erroné dans ce score représente les mots catégorisés comme adjectifs mais qui n'en sont pas et sont absents du le lexique utilisé par *AliAdj* ce qui les rend non traitables par le système. Ils relèvent donc d'une erreur de catégorisation qui a eu lieu lors du traitement des données par *TreeTagger* (cf. partie 3. Quelques erreurs d'AliScol du Chapitre 3. Prétraitement des données, p. 39).

Si même ce pourcentage d'erreur n'est pas du fait d'*AliAdj*, le système paraît trop parfait, il faut donc prendre du recul sur ces résultats et se demander à quoi ils peuvent être dus.

2. Analyse critique des performances d'AliAdj

Tout d'abord, pour que l'évaluation de cet outil soit plus objective, il aurait fallu que le corpus de référence soit annoté par différentes personnes afin de pouvoir calculer le score inter-annotateur.

Par ailleurs, d'après Antoine (2016), on sait que la plupart des systèmes de TAL ont pour but de coller au mieux à un jeu de données extrait du problème. Or, si la représentativité du corpus n'est pas totalement garantie, la significativité des résultats est à

remettre en question. En effet, il y a plusieurs explications possibles à ces résultats très voire trop bons :

- Les formes produites par les élèves présentent une certaine cohérence c'est pourquoi on arrive assez facilement à retrouver les découpages modélisés : les productions ne relèvent pas de l'aléatoire.
- Les données ne sont pas suffisantes du point de vue du nombre et de la variété. Il faudrait tester l'outil sur d'autres niveaux scolaire (ce qui est prévu) ainsi qu'avec d'autres consignes pour essayer d'obtenir une plus grande diversité de formes adjectivales.
- Les données sont simples. Si le système travaillait sur des données plus compliquées d'un point de vue orthographique, comme sur des dissertations par exemple, peut-être que les résultats seraient différents.

Il faudrait donc pouvoir ré-évaluer *AliAdj* à l'aide d'autres jeux de données pour vérifier les performances du système.

Après avoir remis en perspective l'étonnante efficacité du module de traitement des adjectifs il est temps de s'intéresser aux données qu'il a pu produire ainsi qu'à leur(s) interprétation(s).

Partie 4

-

Résultats

Chapitre 7. Analyse des données

Une fois *qu'AliAdj* a pu traiter l'ensemble des données du corpus, il faut analyser les résultats obtenus et voir s'ils se conforment ou non à l'hypothèse de départ selon laquelle la base de l'adjectifs et ses flexions ne présentent pas la même difficulté d'apprentissage chez les élèves.

1. Observations

Après avoir passé l'ensemble des données dans le module de traitement, on obtient un fichier par niveau contenant uniquement les adjectifs des productions associés à leur description en termes de base et de flexions notamment (cf. Tableau 18, p. 55 et Figure 21, p. 61). Pour faciliter l'analyse de ces données, j'ai écrit un algorithme[44] calculant des pourcentages en fonction de ce que l'équipe E-Calm souhaite observer.

Premièrement, j'ai cherché à distinguer les formes normées des différents types d'erreurs décrites dans le Tableau 22 et la Figure 22. La catégorie « Erreurs bases + flexions » regroupe les formes sur lesquelles il y a à la fois une erreur de base et de flexions (formes non comptabilisées dans les catégories « Erreurs bases » et « Erreurs flexions »). Par ailleurs derrière le terme « flexion », ce sont les erreurs de genre et de nombre qui sont regroupées.

Niveau	Formes normées	Erreurs bases	Erreurs flexions	Erreurs bases + flexions	Total
CP	56,99%	20,43%	12,1%	10,48%	100%
CE1	53,24%	27,32%	10,8%	8,64%	100%
CE2	65,44%	16,52%	11,69%	6,35%	100%
CM1	69,85%	13,48%	11,42%	5,25%	100%
3eme	83,65%	6,27%	8,37%	1,71%	100%
licence1	93,94%	3,79%	2,27%	0%	100%

Tableau 22 : Répartition des adjectifs en fonction de si leur base et/ou flexion sont normées ou erronées.

[44] Disponible à telle adresse : https://github.com/RachelGaubil/Outiller_description_morpho_adjectivale_primaire_universite/tree/master/statistiques

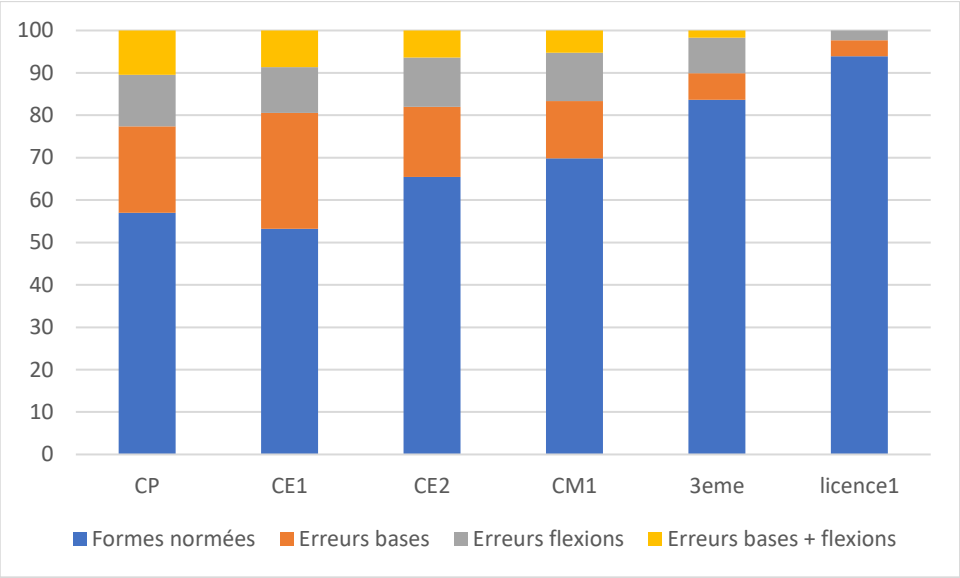

Figure 22 : Répartition des adjectifs en fonction de si leur base et/ou flexion sont normées ou erronées.

On remarque ici que les formes normées sont de plus en plus fréquentes au fur et à mesure des années : 56,99% en CP contre 93,94% en 1ère année de licence, cela malgré une légère diminution de 3,75% entre le CP et le CE1. Cela pourrait être imputé à une augmentation du nombre d'adjectifs utilisés et/ou aux consignes données aux élèves qui diffèrent entre le CP et la période CE1-CM1.

Les erreurs portant sur la base seule des mots sont les plus nombreuses depuis le CP jusqu'au CM1 et vont en diminuant (de 20,43% en CP à 13,48% en CM1). Les erreurs sur les flexions seules semblent elles stagner entre 10,8% et 12,1% du CP au CM1 ; en effet, ce n'est qu'à partir de la 3ème que l'on observe une réelle diminution avec 8,37% d'erreurs et enfin la 1ère année de licence, elle, est à seulement 2,27% d'erreurs.

Cependant, ces chiffres sont à interpréter avec précaution. En effet, il faut également prendre en compte les mots dont l'erreur porte à la fois sur la base et sur la flexion ; ils représentent au maximum 10,48% des adjectifs (en CP). Leur quantité diminue progressivement tout au long des niveaux scolaires, pour disparaître totalement en licence.

J'en conclus donc une tendance globale à la baisse des erreurs peu importe où elles se situent dans le mot. À part dans le supérieur, les difficultés semblent porter aussi bien sur la base que sur les flexions contrairement aux observations menées sur les formes verbales (Wolfarth *et al*., 2018) où les erreurs de flexion restent une difficulté supérieure à celles sur les bases aux différents niveaux de primaire.

Afin de pouvoir mieux observer l'évolution de l'acquisition des bases et des flexions séparément, le Tableau 23 et la Figure 23 présentent les résultats obtenus en opposant les erreurs sur la base du mot et celles sur la flexion du mot

Niveau	Bases normées	Erreurs bases	Total	Flexions normées	Erreurs flexions	Total
CP	69,09%	30,91%	100%	77,42%	22,58%	100%
CE1	64,04%	35,96%	100%	80,56%	19,44%	100%
CE2	77,13%	22,87%	100%	81,96%	18,04%	100%
CM1	81,27%	18,73%	100%	83,32%	16,68%	100%
3eme	92,02%	7,98%	100%	89,92%	10,08%	100%
licence1	96,21%	3,79%	100%	97,73%	2,27%	100%

Tableau 23 : Proportion de bases et de flexions normées ou non

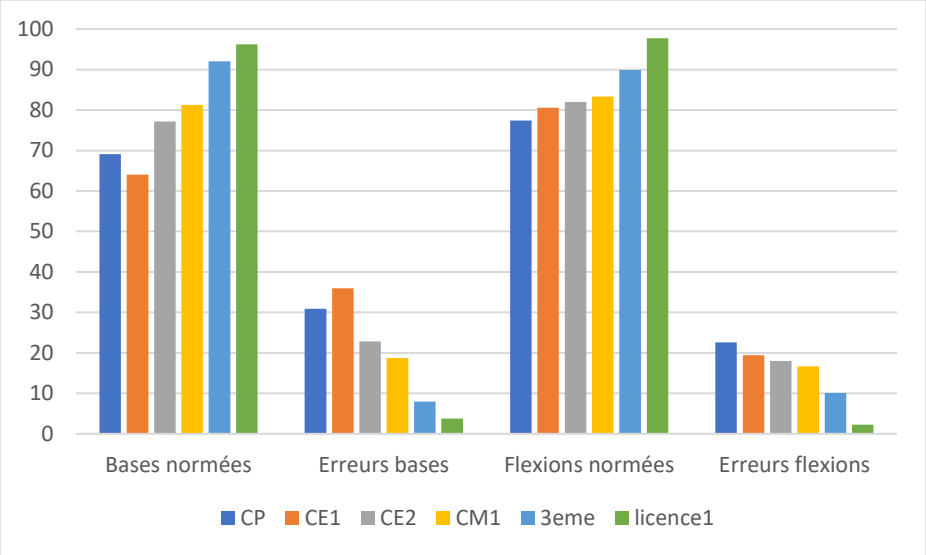

Figure 23 : Proportion de bases et de flexions normées ou non

Ce schéma fait également ressortir une hausse des bases normées au fil du temps, à l'exeption du passage entre le CP (69,09%) et le CE1 (64,04%) où apparaît une légère diminution de ces bases comme observé précédemment.

Il en va de même pour les flexions dont les erreurs sont en constante diminution. Cependant cette décroissance se fait plus marquante du CM1 à la 3ème (moins 6,6%) et de la 3ème à la 1ère année de licence (moins 7,81%) qu'entre les niveaux précédents pour lesquels les différences vont de 3,14% à 1,36%). Ce saut significatif peut s'expliquer par le manque de données intermdiaires, ou par les différentes cohortes d'élèves.

Je souhaite maintenant savoir s'il y existe une différence significative entre cette diminution des erreurs de base et des erreurs de flexions.

Ainsi, dans le Tableau 24 et la Figure 24 ce sont sur les erreurs de base et de flexions que je vais me focaliser.

Niveau	Formes normées	Erreurs bases	Erreurs flexions	Total
CP	45,43%	30,91%	23,66%	100%
CE1	42,82%	35,96%	21,22%	100%
CE2	57,36%	22,87%	19,77%	100%
CM1	63,28%	18,73%	17,99%	100%
3eme	80,99%	7,98%	11,03%	100%
licence1	93,94%	3,79%	2,27%	100%

Tableau 24 : Répartitions des adjectifs en fonction de leur type d'erreur s'il y en a

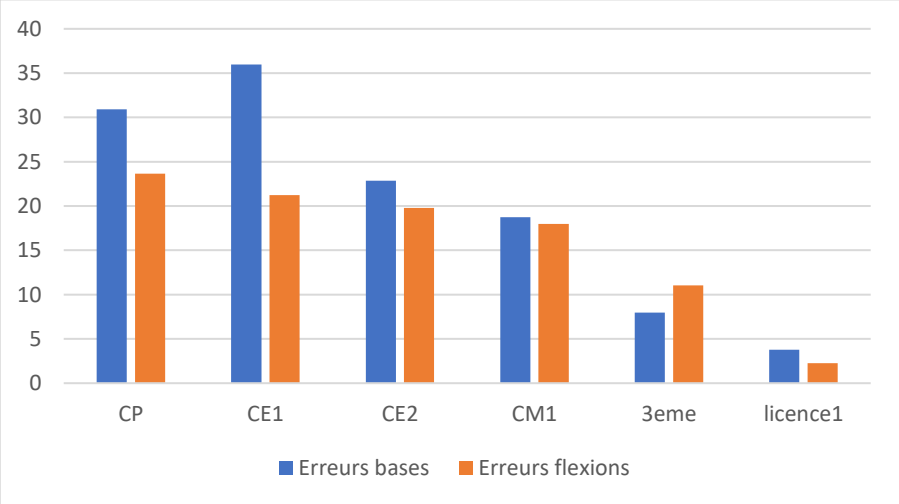

Figure 24 : Répartition des adjectifs erronés en fonction de si l'erreur est sur la base ou la flexion

Comme observé avec la Figure 22, cet histogramme indique que les erreurs de bases sont globalement plus fréquentes que celles de flexions, sauf pour le niveau de 3ème ou l'inverse se produit. On remarque également que la progression concernant les bases est assez nette tandis que celle concernant les flexions se fait plus en douceur.

On remarque toujours une augmentation des erreurs de base entre le CP et le CE1 (plus 5,05%) mais également deux fortes diminutions de ces erreurs : une du CE1 au CE2 (moins 13,09%) et une du CM1 à la 3ème (moins 10,75%).

Pour ce qui est des erreurs de flexions, elles diminuent petit à petit durant l'école primaire (entre 1,45% et 2,44% d'écart entre les niveaux) et un peu plus fortement au

moment du passage au collège et dans le supérieur : 17,99% en CM1 contre 11,03% en 3ème et 2,27% en 1ère année de licence.

Ces constations semblent indiquent que les erreurs de base sont généralement plus fréquentes mais diminuent aussi très rapidement au fil du temps. A l'inverse, le nombre d'erreurs de flexion diminue moins fortement durant l'école primaire et plus fortement dans les niveaux scolaires supérieurs. Cela pourrait laisser penser que les flexions sont une zone de résistance sur laquelle les apprentissages se font moins rapidement. Par ailleurs, la progression en pic sur les flexions est à interpréter avec prudence car le manque de données de niveaux intermédiaires peut y être pour beaucoup.

Afin de pouvoir mieux observer une potentielle différence entre les deux types de flexion soit de genre et de nombre, je les ai dissociées l'une de l'autre (cf. Tableau 25 et Figure 25).

Niveau	Erreurs genre	Erreurs nombre	Total
CP	11,02%	12,63%	23,65%
CE1	9,66%	11,56%	21,22%
CE2	9,02%	10,75%	19,77%
CM1	8,05%	9,94%	17,99%
3eme	5,7%	5,32%	11,02%
licence1	0%	2,27%	2,27%

Tableau 25 : Répartition des adjectifs en fonction du type de flexion erronée

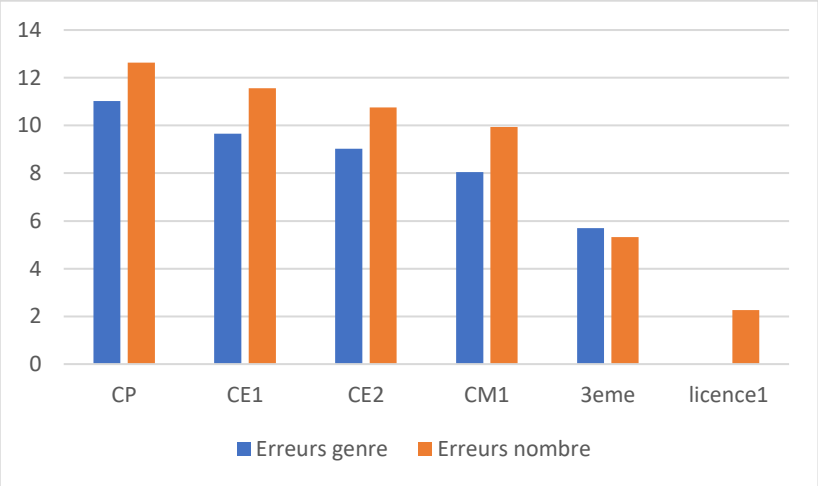

Figure 25: Répartition des adjectifs en fonction du type de flexion erronée

Le graphique montre que les erreurs de nombre sont plus fréquentes à presque tous les niveaux excepté la 3ème mais ces écarts ne dépassent pas les 2,27%. De plus, c'est l'unique type d'erreurs de flexion présent en licence. Cela ne paraît pas très significatif, mais l'interprétation est à modérer en raison du manque de données entre le CM1 et la licence. Par ailleurs, l'information marquante ici, au-delà de la diminution globale des erreurs de flexion est la disparition totale des erreurs de nombre en 1ère année de licence.

L'ensemble des résultats observé montre une progression globale de l'acquisition de l'orthographe des adjectifs malgré quelques points divergents. Dans la partie suivante je vais proposer une interprétation de ces résultats.

2. Interprétation

Les résultats qui viennent d'être présentés montrent que le nombre d'erreurs quelles qu'elles soient diminuent tout au long de la scolarité.

On observe cependant, une légère augmentation des erreurs sur la base des mots entre le CP et le CE1. Cela pourrait s'expliquer par l'expansion du vocabulaire entre ces deux niveaux. Pour déterminer la taille du vocabulaire, j'ai calculé le nombre de lemmes adjectivaux différents les uns des autres par rapport au nombre d'adjectifs total. Le tableau ci-dessous présente donc le pourcentage d'adjectifs différents utilisés à chaque niveau (cf. Tableau 26).

	CP	CE1	CE2	CM1	3eme	Licence1
Taille de vocabulaire	8,21%	17,45%	16,02%	16,43%	30,57%	66,21%

Tableau 26 : Évolution de la taille du vocabulaire en fonction du niveau scolaire

C'est lors du passage du CP au CE1 que l'on observe la plus grosse augmentation de la taille du vocabulaire (plus 9,24%). L'ajout d'autant de nouveaux adjectifs pourrait expliquer l'augmentation du nombres d'erreurs sur la base entre ces deux niveaux. Toutefois, il faut également tenir compte de la différence entre la consigne donnée au CP et celle donnée du CE1 au CM1. En effet, la consigne du CP (cf. partie 1.1 Récolte des données du Chapitre 1. Contexte et problématique, p.12) n'est pas très propice à l'emploi d'adjectifs variés. Par

ailleurs, on remarque que les niveaux avec une consigne commune (du CE1 au CM1) ont une taille de vocabulaire à peu près constante (entre 16,02% et 17,45%).

L'augmentation en flèche de la taille du vocabulaire dans les niveaux supérieurs pourrait elle aussi être due aux consignes. En effet, la consigne donnée aux élèves de 3ème diffère de celles données en primaire. Par ailleurs, les écrits de licences ne résultent d'aucune consigne de chercheur puisque ce sont des productions élaborées dans le cadre des cours. Cela explique sûrement en partie la taille du vocabulaire plus élevée.

Concernant la différence d'acquisition entre la base d'un mot et sa(ses) flexion(s), les graphiques et tableaux précédents montrent que c'est la base qui est sujette au plus grand nombre d'erreurs. Les deux ne s'acquièrent donc pas de la même manière. Cependant, c'est également sur la base que l'on observe une amélioration des résultats plus rapide. Peut-être cela signifie-t-il que l'apprentissage des flexions constitue une zone de résistance plus importante que celle du lexique ?

A propos des flexions de genre et de nombre, une certaine similitude de progression est observée. En effet, les proportions des erreurs qui portent sur ces flexions diminuent tout au long de la scolarité et ne s'écartent pas l'une de l'autre de plus de 2,27% à chaque niveau. Cependant les erreurs sur les flexions de nombre restent légèrement plus nombreuses tout au long de la scolarité (sauf en 3ème) et elles disparaissent totalement au niveau de la licence. Cela pourrait être dû au fait que le genre est plus souvent marqué à l'oral que le nombre et donc plus facile à maitriser car c'est sur les variations phonologiques que les élèves s'appuient (David & Wattelet, 2016 ; Cogis & Brissaud, 2019).

Par ailleurs, la tendance qui fait que les erreurs de bases sont plus fréquentes que les erreurs de flexions semblent s'inverser après l'école primaire. Cependant, je ne dispose pas d'assez de données pour avoir un regard objectif à ce propos. Il serait donc intéressant de pouvoir ajouter à ce travail des données des niveaux intermédiaires entre le CM1 et la 1ère année de licence ce qui est prévu dans le projet.

De plus la progression rapide qui est montrée au niveau de la 3ème et de la licence peut aussi s'expliquer par le changement des groupes d'élèves suivis. En effet, du CP au CM1, c'est une étude longitudinale qui a pu être mise en place et ce sont donc les mêmes

élèves avec la même consigne (du CE1 au CM1) qui ont été suivis à travers le temps, ce qui n'est pas le cas pour les niveaux scolaires supérieurs.

En conclusion, le nombre d'erreurs portant sur la base des adjectifs est globalement plus élevé partout mais suit une décroissance plus importante. Les erreurs de flexions sont donc moins nombreuses mais mettent plus de temps à diminuer.

Cela nous conforte dans l'idée que la base et les flexions de l'adjectifs ne s'acquierent pas à la même vitesse et ne présentent donc pas la même difficulté pour les élèves. Néanmoins, on ne peut pas dire que les erreurs de bases disparaissent au profit des erreurs de flexions. Les deux restent présentes tout au long de la scolarité et n'ont simplement pas la même vitesse de diminution.

Une interprétation possible à ce phénomène serait que les flexions de genre comme de nombre constituent une zone de résistance au niveau de laquelle l'apprentissage se fait plus lentement, comme ce qui a été observé pour les verbes par Wolfarth *et al.* (2018).

Par ailleurs, il est également à noter que la flexion de nombre semble poser plus de difficultés à être maitrisée que la flexion de genre même si la différence entre les deux n'excède pas 2,27%.

Conclusion et perspectives

1. Conclusion

Dans ce projet, les comportements morphologiques des adjectifs ont pu être décrits selon un modèle spécifique à cette recherche grâce à une liste de comportements découpant chaque mot en trois parties : *base – flexion de genre – flexion de nombre*. Ces représentations ont permis la conception d'un outil pouvant faire ces découpages de façon automatique sur des formes normées comme sur des formes erronées. Le système fait également la comparaison entre les deux et en déduit le type d'erreur que porte le mot s'il y en a.

Les résultats extraits montrent une progression orthographique sur les adjectifs tout au long de la scolarité. Cette amélioration est plus forte sur les bases des mots, qui par ailleurs ont un taux d'erreurs plus élevé à presque chaque niveau. Cela montre bien que l'acquisition orthographique de la base des adjectifs n'est pas la même que celle de ses flexions qui paraissent opposer une certaine résistance par la faible progression observée. Ces constatations vont dans le sens d'un apprentissage parallèle entre les morphologies verbale et adjectivale. Par ailleurs, une légère différence a été observée entre l'évolution des erreurs sur les flexions de genre et sur les flexions de nombre. Ces dernières sont en effet plus présentes, et ce, à chaque niveau scolaire observé.

Grâce à tout ceci, l'hypothèse de départ selon laquelle l'acquisition de l'orthographe adjectivale est un apprentissage à deux vitesses (opposant base et flexions) semble pertinente mais reste à confirmer sur un jeu de données plus large et plus diversifié. Par ailleurs, les objectifs sont remplis puisque l'outil développé dans ce travail propose une description linguistique des adjectifs produits par des apprenants. Il permet également de se rendre compte de la localisation des erreurs ainsi que de leur évolution au cours du temps.

2. Perspectives

Pour compléter le travail effectué ici, il faudrait pouvoir relancer *AliAdj* en ayant préalablement retiré du corpus Resolco les phrases correspondant aux consignes, soit aux bandelettes de papier collées dans le texte. En effet, l'orthographe correcte de ces mots

n’est absolument pas représentative des capacités des élèves. Par ailleurs, il serait également judicieux de trouver comment analyser les formes hypo-segmentées et hyper-segmentées qui n’ont pas été traitées ici.

Lors de travaux futurs, il serait intéressant de pouvoir examiner plus en détail l’impact de la phonologie sur l’apprentissage morphologique des adjectifs. En effet, on sait que les élèves s’appuient dessus pour apprendre (David & Wattelet, 2016 ; Cogis & Brissaud, 2019).

Il serait également possible d’affiner le traitement des adjectifs en proposant une modélisation qui tiendrait compte du phénomène de surgénéralisation. Il dissocierait donc l’erreur de flexion de nombre en deux : l’absence de marque contre la présence d’une marque erronée. Par exemple, lorsque l’élève écrit « égals », c’est sûrement qu’il a retenu le *-s* comme flexion de nombre et qu’il généralise ce modèle au mot « égal ». Il serait donc indiqué dans ce cas que le nombre est respecté mais pas la marque de nombre.

Et enfin, une analyse des erreurs de flexion de genre en fonction du genre des scripteurs pourrait également être pertinente puisqu’on sait que les élèves ont tendance à confondre genre grammatical et genre narratif (Cogis, 2003 ; Cogis & Brissaud, 2019).

3. Bilan personnel

Le travail effectué durant ce stage m’a permis de mieux me rendre compte de ce qu’est réellement le travail de recherche : beaucoup de documentation, beaucoup de questions et peu de théories unanimes. Cependant, c’est aussi un travail riche de collaborations car il est important d’avoir un regard neuf et des avis divergents pour pouvoir avancer objectivement ; c’est lors des réunions avec les membres du projet *E-Calm* que j’ai pu m’en rendre compte.

Pendant ce stage, j’ai pu mettre à profit les enseignements de ce master sur le plan théorique comme sur le plan pratique avec tout ce qui concerne la modélisation linguistique et la conception de système de traitement automatique du langage. De plus, appliquer concrètement tout ce que j’ai appris ces deux dernières années permet de se rendre compte de tout l’utilité du TAL dans le domaine de la recherche.

Bien que la majorité de ce stage ait été effectuée en télétravail à cause du confinement dû à la crise sanitaire, mon travail n’a pas réellement été entravé. En effet, beaucoup de ressources sont disponibles en ligne de nos jours et M. Ponton, mon encadrant,

a été présent à chaque fois que j'ai pu rencontrer des difficultés. Les obstacles principaux auxquels j'ai pu me heurter sont la gestion du temps lors du télétravail et les choix linguistiques à faire puisqu'il n'existe aucune bonne réponse : tout est affaire de point de vue et d'objectif. C'est ce que je retiendrai de ce projet, ainsi que l'importance de la collaboration au sein d'une équipe.

Bibliographie

Antoine, J.-Y. (2016, 17 octobre). Évaluation en Traitement Automatique des Langues : Rigueur scientifique, course d'un jour ou aveuglement collectif ? *Ethique et TAL*. https://www.ethique-et-tal.org/2016/10/17/evaluation-en-traitement-automatique-des-langues-rigueur-scientifique-course-dun-jour-ou-aveuglement-collectif/

Blanchard, A. (2006). *Analyse morphologique des réponses d'apprenants en environnement d'Apprentissage Assisté par Ordinateur*. [Mémoire de master, Université Stendhal-Grenoble III]. http://blanchard.alexia.free.fr/doc_site/memoire.pdf

Catach N. (1980, 3ème édition, 1995). L'orthographe française : traité théorique et pratique avec des travaux d'application et leurs corrigés (avec la collaboration de Claude Gruaz et Daniel Duprez). Paris : Nathan

Cogis, D. (2003). *Marques orthographiques du féminin et pratiques de l'écrit*. *27*, 103-115.

Cogis, D., & Brissaud, C. (2019). A la poursuite des marques de genre... In C. Mortamet, *L'orthographe : Pratiques d'élèves, pratiques d'enseignants, représentations* (p. 43-71). Presses universitaires de Rouen et du Havre.

David, J., & Wattelet, S. (2016). Approcher, comprendre, maitriser l'orthographe grammaticale au CE1. *Le français aujourd'hui*, *192*(1), 73. https://doi.org/10.3917/lfa.192.0073

Fradin, B. (1994). L'approche à deux niveaux en morphologie computationnelle et les développements récents de la morphologie. *Morphologie computationnelle*, *35*(2), 9-48.

Fuchs C., 1993, Danlos L., Lacheret-Dujour A., Luzatti D., Victorri B., Linguistique et Traitements Automatiques des Langues. Hachette, France, 1993.

Garcia-Debanc, C. (2018). Ajout et résolution de problèmes de cohésion textuelle : Analyses linguistiques de textes d'élèves et présentation de différents dispositifs de travail pour enseigner l'ajout au cycle 3. *Repères*, *57*, 185-208.

Goigoux, R. (2015). *Lire et Écrire. Étude de l'influence des pratiques d'enseignement la lecture et de l'écriture sur la qualité des premiers apprentissages*. Institut Français de l'Éducation. Rapport de recherche.

Gourdet, P. (2017). Les exercices à l'école élémentaire et l'apprentissage de la langue : Quelle(s) réalité(s) ? *Repères. Recherches en didactique du français langue maternelle*, *56*, 51-72. https://doi.org/10.4000/reperes.1193

Lallich-Boidin, G. (1987). *Traitement automatique de la langue naturelle : L'analyse morphologique flexionnelle du français écrit*. *87-3*.

Laparra, M. (2010). Pour un enseignement progressif de l'orthographe dite grammaticale du français. Comment résoudre les contradictions entre les besoins discursifs des scripteurs et les contraintes imposées par le fonctionnement de la langue ? *Repères. Recherches en didactique du français langue maternelle*, *41*, 35-46. https://doi.org/10.4000/reperes.278

Lefrançois, P. (2009). Évolution de la conception du pluriel des noms, des adjectifs et des verbes chez les élèves du primaire. *Repères. Recherches en didactique du français langue maternelle*, *39*, Article 39. https://doi.org/10.4000/reperes.846

Noailly, M. (1999). *L'adjectif en français* (Ophrys).

OCDE (2019). *Résultats du PISA 2018 (Volume I) : Savoirs et savoir-faire des élèves*. PISA. https://doi.org/10.1787/ec30bc50-fr

OCDE, & Statistique Canada. (2000). *La littératie à l'ère de l'information*. http://www.oecd.org/fr/education/innovation-education/39438013.pdf

Popescu-Belis, A. (2007). Le rôle des métriques d'évaluation dans le processus de recherche en TAL. *TAL, 48, 67-91*.

Roché, M. (2010). Base, thème, radical. *Recherches linguistiques de Vincennes*, *39*, 95-134. https://doi.org/10.4000/rlv.1850

Séguin, H. (1973). *Le genre des adjectifs en français, analyse quantitative et correspondances phonographiques des règles*. *20*, 52-74.

Silberztein, M. (1996). Analyse automatique de corpus avec INTEX. *Lexique, syntaxe… automatique. Hommage à Jean Dubois*, *LINX*, *34-35*, 269-276. https://doi.org/10.3406/linx.1996.1435

Wehrli E. (1997). L'analyse syntaxique des Langues Naturelles. Problèmes et Méthodes. Masson, Paris.

Wolfarth, C. (2019). *Apport du TAL à l'exploitation linguistique d'un corpus scolaire longitudinal*. [Thèse de doctorat, Université Grenoble Alpes]. tel.archives-ouvertes.fr. https://tel.archives-ouvertes.fr/tel-02517320/document

Wolfarth, C., Ponton, C., & Brissaud, C. (2018). Gestion de la morphographie verbale en production d'écrits : Que peut nous apprendre un corpus longitudinal ? *Repères. Recherches en didactique du français langue maternelle*, *57*, 209-226. https://doi.org/10.4000/reperes.1576

Wolfarth, C., Brissaud, C., & Ponton, C. (2018). Transcrire et normer un corpus scolaire : Pour quelles analyses ? *Diptyque*, *36*, 121-145.

Sitographie

ABU - Listes de mots communs. (s. d.). http://abu.cnam.fr/DICO/mots-communs.html [dernière consultation le 10/03/2020]

BASE : Définition de BASE. (s. d.). CNRTL. https://www.cnrtl.fr/definition/base [dernière consultation le 29/08/2020]

Béchet, F. (2001). LIA-PHON: Un système complet de phonétisation de textes. TAL. Traitement automatique des langues, 42(1), 47-67. http://pageperso.lif.univ-mrs.fr/~frederic.bechet/download.html [consulté le 24/04/2020]

FLEXION : Définition de FLEXION. (s. d.). CNRTL. https://www.cnrtl.fr/definition/flexion [dernière consultation le 29/08/2020]

Frederic Bechet. (s. d.). http://pageperso.lif.univ-mrs.fr/~frederic.bechet/download.html [dernière consultation le 24/04/2020]

GRAPHÈME : Définition de GRAPHÈME. (s. d.). CNRTL. https://www.cnrtl.fr/definition/graph%C3%A8me [dernière consultation le 22/07/2020]

E-CALM. (s. d.). E-CALM. http://e-calm.huma.num.fr [dernière consultation le 10/07/2020].

E-CALM | ORTOLANG. (s. d.). https://www.ortolang.fr/market/corpora/e-calm [dernière consultation le 23/08/2020]

Helmut S. (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods.Language Processing, Manchester, UK. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ [dernière consultation le 24/04/2020].

LEXÈME : Définition de LEXÈME. (s. d.). https://www.cnrtl.fr/definition/lex%C3%A8me [dernière consultation le 16/01/2020]

lemme—Définitions, synonymes, conjugaison, exemples | Dico en ligne Le Robert. (s. d.). https://dictionnaire.lerobert.com/definition/lemme [dernière consultation le 02/09/2020]

LIDILEM. UGA. (s. d.). Scoledition. http://scoledit.org/scoledition/index.php [dernière consultation le 24/08/2020]

MORPHÈME : Définition de MORPHÈME. (s. d.). https://www.cnrtl.fr/lexicographie/morph%C3%A8me [dernière consultation le 16/01/2020]

New, B., & Pallier, C. (1999). Lexique. http://www.lexique.org/ [dernière consultation le 10/03/2020]

Wikipedia. (s. d.). https://www.wikipedia.org/ [dernière consultation le 10/03/2020]

Glossaire

Graphème : ensemble minimal de lettres transcrivant un phonème ou ayant une fonction morphologique ou étymologique(https://www.cnrtl.fr/definition/graph%C3%A8me).

Lemme : c'est la forme canonique d'un mot, généralement le masculin singulier ou l'infinitif pour un verbe (https://dictionnaire.lerobert.com/definition/lemme).

Lexème : morphème lexical, soit l'unité minimale de signification appartenant au lexique (https://www.cnrtl.fr/definition/lex%C3%A8me) ; tandis qu'un morphème est la plus petite unité de sens et n'est pas forcément lexical.

Littératie : définie par l'OCDE[45] comme « l'aptitude à comprendre et à utiliser l'information écrite dans la vie courante, à la maison, au travail et dans la collectivité en vue d'atteindre des buts personnels et d'étendre ses connaissances et ses capacités » (OCDE & Statistique Canada, 2000, p10).

Morphème : unité minimale de signification.

Token : de l'anglais « jeton », utilisé dans le domaine de la linguistique il désigne une unité lexicale.

[45] Organisation de Coopération et de Développement Économiques

Sigles et abréviations utilisés

ANR : Agence Nationale de la Recherche

CIRCEFT-ESCOL : Centre Interdisciplinaire de Recherche « Culture, Éducation, Formation, Travail » - Éducation et Scolarisation

CLESTHIA : CLESTHIA – Langage, systèmes, discours

CLLE-ERSS : Cognition, Langues, Langage, Ergonomie – Équipe de Recherche en Syntaxe et Sémantique

CRISS : Centre de Recherche en Informatique appliquée aux Sciences Sociales

DGESCO : Direction Générale de l'Enseignement Scolaire

E-CALM : Ecriture scolaire et universitaire : Corpus, Analyses Linguistiques, Modélisations didactiques

LIDILEM : Linguistique et Didactique des Langues Étrangères et Maternelles

OCDE : Organisation de Coopération et Développement Économiques

PISA : Programme International pour le Suivi des Acquis des élèves

TAL : Traitement Automatique du Langage

Table des illustrations

Table des tableaux

Table des annexes

Annexe 1
Tâches du projet E-Calm

Tableau récapitulatif des différentes tâches du projet E-Calm.

Tâche	**Laboratoire responsable**	**Objectif** et sous-tâches
0. Coordination	CLESTHIA	**Superviser le projet** : il faut mettre en relation les équipes et vérifier l'avancement des tâches ainsi que le respect du planning et du budget. Faire une veille scientifique régulière est nécessaire, tout comme surveiller la bonne coarticulation des tâches entre elles. Ce laboratoire s'occupe aussi d'organiser les réunions de travail et la communication du projet à l'intérieur et à l'extérieur des équipes.
1. Structuration d'un corpus cohérent et significatif	LIDILEM	**Produire un corpus XML-TEI** : il faut créer le corpus brut (sélection des données, transcriptions, transposition dans le bon format, etc.), compléter les données si nécessaire et sélectionner les métadonnées intéressantes dans ce projet. Il faut également choisir une norme de transcription et définir un schéma d'encodage pour pouvoir sortir le corpus sous la norme TEI.
2. Analyse des écrits scolaires et académiques : orthographe grammaticale et lexicale	LIDILEM	**Montrer les performances orthographiques au niveau de la morphosyntaxe verbale, des accords adjectivaux et des morphèmes dérivatifs** : pour cela on se base sur une étude longitudinale des élèves de primaire (cf : corpus *Scoledit*), une étude transversale (niveaux primaire, secondaire et universitaire) et une étude de toutes les réécritures faites par les élèves dont celles induites par les enseignants. Les entretiens métagraphiques (avec des élèves de CE1, 5° et 1ère année de licence).
3. Analyse des écrits scolaires et	CLLE	**Produire des guides d'annotation et des indicateurs de compétences** : cela commence par l'annotation du sous corpus *Resolco* dans le but d'analyser la cohésion et la cohérence. À partir de l'analyse précédente, il faut créer des d'indicateurs de compétences discursives des élèves pour pouvoir comparer les

académiques : cohérence discursive		différences entre genres textuels d'un même niveau scolaire et pour pouvoir voir la progression durant le travail d'écriture avec une comparaison entre les brouillons et les productions finales.
4. Analyse des interventions écrites des enseignants sur les écrits des élèves	CIRCEFT	**Typologie des interventions et analyse des corrélations entre celles-ci** : il faut créer une typologie des interventions : une description d'où (orthographe, cohésion, etc.), quand (brouillon, étape intermédiaire, etc.) et comment (soulignements, commentaires, etc.) sont faites ces interventions sur les copies. Ensuite lors d'un entretien, il est demandé aux apprenants de commenter les interventions et leur réaction face à celles-ci. Enfin, il y a une analyse des révisions qu'on faites les élèves suite aux interventions des enseignants
5. Résultats et interprétation	CIRCEFT	**Synthèse des tâches 2, 3, 4 et 6 et observation des corrélations entre compétence orthographique, discursive et intervention des enseignants** : le but ici est la mise en relation des annotations (de tout le projet) et de leurs interprétations. Il faut également s'assurer de la cohérence des différents guides d'annotation qui ont été produits dans les étapes précédentes, soit vérifier qu'ils peuvent être généralisés à l'ensemble du corpus. Enfin, les corrélations analysées sont mises en relation avec les contextes sociologiques et didactiques.
6. Exploitations du corpus à des fins de formation et d'enseignement	CLLE	**Synthèse des indicateurs pour l'orthographe et la cohérence des textes et modules de formation aux enseignants :** il faut procurer aux enseignants les indicateurs de compétences (orthographiques et de cohérence) et de progrès construits durant ce projet et les aider enseignants à se former à l'évaluation des textes d'élèves en utilisant un échantillon du corpus. Il s'agit aussi de leur fournir des échantillons de textes avec leurs consignes pour leur permettre de faire des activités d'évaluation et de réécriture avec leurs élèves.
7. Diffusion et valorisation des ressources produites	CLESTHIA	**Créer le site Web contenant l'ensemble des ressources du projet** : il faut donc construire et compléter au fur et à mesure le site internet : mettre en ligne de l'ensemble des ressources, des guides et des outils pour le traitement informatique des données conçus tout au long du projet. Il faudra finir par une évaluation du site internet et, si nécessaire, son adaptation.

Annexe 2
Données du projet E-Calm

Tableau récapitulatif des données du projet *E-Calm* avec un total de 6 754 textes soit environ 4 937 700 mots.

	Nom du corpus	Concepteur	disponibilité	transcrit ou non	Genre	Plusieurs versions	Intervention enseignant	Taille moyenne / texte	Nb textes	Taille moyenne/texte en nb mots	Total mots
CP	Scoledit	LIDILEM	oui	oui	narratif	non	non	4 phrases	965	30	28950
CE1	Ecriscol	CLESTHIA	déc-17	en partie	narratif	oui	oui	paragraphe	175	80	14000
	Scoledit	LIDILEM	oui	oui	narratif	non	non	paragraphe	800	80	64000
CE2	Scoledit	LIDILEM	déc-17	non	narratif	non	non	page	800	500	400000
	Ecriscol	CLESTHIA	juin-17	non	narratif	non	oui	1/2 page	50	250	12500
	Resolco	CLLE	déc-18	non	narratif	non	en partie	1/2 à 1 page	75	300	22500
CM1	Scoledit	LIDILEM	avr-18	non	narratif	non	non	page	800	500	400000
	Resolco	CLLE	déc-18	non	narratif	non	parfois	1 page	75	500	37500
CM2	Scoledit	LIDILEM	déc-18	non	narratif	non	non	page	800	500	400000
	Ecriscol	CLESTHIA	déc-17	en partie	narratif	oui	oui	page	575	500	287500
	Resolco	CLLE	déc-18	non	narratif	non	parfois	1 page	75	500	37500
6e	Ecriscol	CLESTHIA	déc-17	non	narratif	non	non	page	120	500	60000
	Resolco	CLLE	déc-18	non	narratif	non	parfois	1à2 pages	75	750	56250
5e	Resolco	CLLE	déc-18	non	narratif	non	parfois	1à2 pages	75	750	56250
4e	Ecriscol	CLESTHIA	juin-17	non	narratif	non	non	2pages	30	1000	30000
	Resolco	CLLE	déc-18	non	narratif	non	parfois	1à2 pages	75	750	56250
3e	Ecriscol	CLESTHIA	déc-17	non	narratif	oui	oui	2pages	90	1000	90000
	Resolco	CLLE	déc-18	non	narratif	non	parfois	1à2 pages	75	750	56250

2nde	Ecriscol	CLESTHIA	juin-17	oui	narratif	non	non	3 pages	70	1500	105000
1ere											
Terminale											
Licence	Ecriscol	CLESTHIA	déc-17	oui	narratif	non	non	1/2 page	450	250	112500
	Littéracie avancée	LIDILEM	oui	oui	argumentatif		en partie	2800 mots	10	2800	28000
	Littéracie avancée	LIDILEM	oui	oui	rapport de stage		en partie	4600 mots	15	4600	69000
	Resolco	CLLE	déc-18	non	narratif	non	parfois	1à2 pages	75	750	56250
Master	Littéracie avancée	LIDILEM	oui	oui	argumentatif		non	3 pages	166	1500	249000
	Littéracie avancée	LIDILEM	oui	oui	argumentatif	en partie	non	1 page	62	500	31000
	Littéracie avancée	LIDILEM	oui	oui	syntheses		en partie	1500 mots	10	1500	15000
	Littéracie avancée	LIDILEM	oui	oui	mémoires		en partie	30 pages	76	15000	1140000
	Resolco	CLLE	déc-18	non	narratif	non	parfois	1à2 pages	50	250	12500
	Mémoires Escol	ESCOL	déc-17	oui	mémoires	oui	non	50 pages	40	25000	1000000

Annexe 3
Prétraitement graphique

Ce prétraitement graphique a été effectué dans le but de régulariser le texte avant de pouvoir le soumettre à l'analyseur morphologique (Antoniadis, 1984, cité par Lallich-Boidin, 1987).

Ainsi, on observe les transformations suivantes :

- Toute majuscule autre que la première lettre d'un nom propre est remplacée par la lettre minuscule correspondante.
- Tout symbole de ponctuation est précédé et suivit d'un espace.
- Toute forme élidée est remplacée par sa forme complète (exemple : *l'* s'écrit *le* ou *la* en fonction du contexte).
- Certaines graphies sont remplacées de la façon suivante :

Graphie originale	ll	nn	ss	tt	ch	u¨	gue	gui	gua	guo	v	gea	geo	ça	ço
Nouvelle graphie	l*	n*	s*	t*	c*	u*	g*e	g*i	g*a	g*o	f*	g_a	g_o	c_a	c_o

- Certaines formes sont éclatées. La liste complète se trouve dans (Berrendonner, 1983, cité par Lallich-Boidin, 1987). Il n'est décrit ici que les éclatements des formes les plus courantes[46] :

  au → à + le
  du → de + le
  aucun → pas + un
  quel → que + %
  dont → que + de + %
  celui → cet + lui

  Le symbole % représente un pronom anaphorique.

[46] Voir Lallich-Bodin (1987, p. 4) pour plus de précision sur les catégories syntaxiques des éclatements.

Annexe 4
Sortie de l'outil *Aliscol*

Voici un exemple de résultat de l'outil *AliScol* lancé sur une production d'un élève de CP.

Id Prod	Id Eleve	Id Classe	Niv	Long Prod	IdTok Norm	Lemme	Catégorie	Seg Norm	Seg Trans	IdTok Trans	IdSeg Trans	Statut Erreur	Statut Segm
815_CP	815		CP	36	1	<s>	[ZTRM->EXCEPTION]	<s>	<s>	1	1	01-Normé	01-Normé
815_CP	815		CP	36	2	un	DET:ART	Un	un	2	1	01-Normé	01-Normé
815_CP	815		CP	36	3	petit	ADJ	petit	petit	3	1	01-Normé	01-Normé
815_CP	815		CP	36	4	chat	NOM	chat	cha	4	1	02-Phono	01-Normé
815_CP	815		CP	36	5	tomber	VER:pres	tombe	donbe	5	1	04-ApproxGraphique	01-Normé
815_CP	815		CP	36	6	de	PRP	d'	de	6	1	04-ApproxGraphique	01-Normé
815_CP	815		CP	36	7	un	DET:ART	une	une	7	1	01-Normé	01-Normé
815_CP	815		CP	36	8	marche	NOM	marche	marque	8	1	04-ApproxGraphique	01-Normé
815_CP	815		CP	36	9	et	KON	et	é	9	1	02-Phono	01-Normé
815_CP	815		CP	36	10	se	PRO:PER	se	cefaimale	10	1	02-Phono	03-HypoSeg
815_CP	815		CP	36	11	faire	VER:pres	fait	cefaimale	10	2	02-Phono	03-HypoSeg
815_CP	815		CP	36	12	mal	ADV	mal	cefaimale	10	3	02-Phono	03-HypoSeg

Annexe 5
Sortie du module d'enrichissement

Voici un exemple de résultat du module d'enrichissement : le même tableau que l'annexe précédente mais auquel trois colonnes ont été ajoutées.

Id Prod	Id Eleve	Id Classe	Niv	Long Prod	IdTok Norm	Lemme	Catégorie	Seg Norm	Seg Trans	IdTok Trans	IdSeg Trans	StatutErreur	Statut Segm	Genre	Nombre	Infover
815_CP	815		CP	36	1	<s>	[ZTRM->EXCEPTION]	<s>	<s>	1	1	01-Normé	01-Normé	–	–	–
815_CP	815		CP	36	2	un	DET:ART	Un	un	2	1	01-Normé	01-Normé	m	s	–
815_CP	815		CP	36	3	petit	ADJ	petit	petit	3	1	01-Normé	01-Normé	m	s	–
815_CP	815		CP	36	4	chat	NOM	chat	cha	4	1	02-Phono	01-Normé	m	s	–
815_CP	815		CP	36	5	tomber	VER:pres	tombe	donbe	5	1	04-ApproxGraphique	01-Normé	–	–	imp:pre:2s;ind:pre:1s;ind:pre:3s;sub:pre:3s;
815_CP	815		CP	36	6	de	PRP	d'	de	6	1	04-ApproxGraphique	01-Normé	–	–	–
815_CP	815		CP	36	7	un	DET:ART	une	une	7	1	01-Normé	01-Normé	f	s	–
815_CP	815		CP	36	8	marche	NOM	marche	marque	8	1	04-ApproxGraphique	01-Normé	f	s	–
815_CP	815		CP	36	9	et	KON	et	é	9	1	02-Phono	01-Normé	–	–	–
815_CP	815		CP	36	10	se	PRO:PER	se	cefaimale	10	1	02-Phono	03-HypoSeg	–	–	–
815_CP	815		CP	36	11	faire	VER:pres	fait	cefaimale	10	2	02-Phono	03-HypoSeg	m	s	ind:pre:3s;par:pas;par:pas;
815_CP	815		CP	36	12	mal	ADV	mal	cefaimale	10	3	02-Phono	03-HypoSeg	–	–	–

Annexe 6
Modèles des comportements adjectivaux

Liste des 29 modèles construits lors de la modélisation linguistique.

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
ancien	ancien + _ + _	ancien + _ + s	ancienn + e + _	ancienn + e + s
attentif	attentif + _ + _	attentif + _ + s	attentiv + e + _	attentiv + e + s
bas	bas + _ + _		bass + e + _	bass + e + s
beau	beau + _ + _	beau + _ + x	bell + e + _	bell + e + s
bée			bée + _ + _	bée + _ + s
blanc	blanc + _ + _	blanc + _ + s	blanch + e + _	blanch + e + s
bonhomme	bonhomme + _ + _	bonhomme + _ + s		
bref	bref + _ + _	bref + _ + s	brèv + e + _	brèv + 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	égal + 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
furax	furax + _ + _			
gris	gris + _ + _		gris + e + _	gris + e + s
inventeur	inventeur + _ + _	inventeur + _ + s	inventric + e + _	inventric + e + s
long	long + _ + _	long + _ + s	longu + e + _	longu + e + s
malin	malin + _ + _	malin + _ + s	malign + e + _	malign + e + s
menteur	menteur + _ + _	menteur + _ + s	menteus + e + _	menteus + e + s
muet	muet + _ + _	muet + _ + s	muett + e + _	muett + e + s
sec	sec + _ + _	sec + _ + s	sèch + e + _	sèch + e + s
vengeur	vengeur + _ + _	vengeur + _ + s	vengeress + e + _	vengeress + e + s
vieux	vieux + _ + _		vieill + e + _	vieill + e + s

Table des matières

MOTS-CLÉS : orthographe, morphologie flexionnelle, modélisation linguistique, TAL (Traitement Automatique du Langage)

RÉSUMÉ

En ce qui concerne les compétences orthographiques des élèves, il n'existe pas de réelle description à cause du manque de données. Le projet E-Calm dans lequel s'inscrit ce mémoire cherche donc à rassembler un ensemble de ressources pour chaque niveau scolaire afin de pouvoir se lancer dans des analyses linguistiques à des fins didactiques, et grâce à la mise en place d'outils de TAL. En effet, la quantité de données récoltée ne permet pas les analyses manuelles. Le travail décrit ici est centré sur les aptitudes des apprenants concernant la morphologie adjectivale.

Dans un premier temps, ce mémoire propose une modélisation linguistique du comportement morphologique flexionnel de l'adjectif en se basant sur des productions d'élèves et d'étudiants. Dans un second temps, il s'intéresse à la conception d'un outil permettant une réalisation automatique de cette modélisation et enfin à l'analyse des résultats du système présenté.

KEYWORDS: spelling, inflectional morphology, linguistic modeling, NLP (Natural Language Processing)

ABSTRACT

Concerning french pupils an students' spelling skills, there is no real description due to a lack of data. The E-Calm project in which this dissertation takes part, therefore, seeks to bring together a set of resources for each school level in order to be able to undertake linguistic analyses for teaching purposes. This is made possible thanks to the implementation of NLP tools; indeed, the amount of data collected does not allow manual analyses. The work described here is focused on the learners' abilities concerning adjectival morphology.

Initially, this masters' thesis proposes a linguistic modeling of the adjective's inflectional morphological behavior based on students' productions. In a second step, it focuses on the design of a tool allowing an automatic realization of this modeling and finally on the analysis of the results of the presented system.