Apport du TAL à la constitution et l'exploitation d'un corpus scolaire de cours préparatoire

Claire Wolfarth

Apport du TAL à la constitution et l'exploitation d'un corpus scolaire de cours préparatoire

Nom : WOLFARTH
Prénom : Claire

UFR LLASIC

Mémoire de master 2 recherche – Sciences du langage

Spécialité ou Parcours : Industries de la langue

Sous la direction de Claude Ponton et Corinne Totereau

TABLE DES MATIÈRES

REMERCIEMENTS

Je voudrais adresser mes remerciements les plus sincères à mes deux encadrants de mémoire, M. Claude Ponton et Mme Corinne Totereau, qui m'ont accompagnée tout au long de ce mémoire, et même jusqu'à Paris. Je les remercie pour leur aide, leurs conseils et leur patience et pour les opportunités qu'ils m'ont données la chance de saisir.

Je souhaiterais également remercier Mme Catherine Brissaud et M. Olivier Kraif qui ont accepté de faire partie de mon jury.

Un grand merci à Laura, collègue de stage et de mémoire, qui n'a pas hésité à me transmettre ses connaissances sur l'orthographe, ses références bibliographiques, ses expériences, ses avancées et ses doutes. Sans oublier que c'est ensemble que nous avons transcrit le corpus utilisé dans ce mémoire.

Je ne souhaiterais pas non plus oublier toutes les personnes qui ont contribuées au recueil du corpus sur lequel repose ce travail et qui ont ainsi contribué à poser la première pierre de mon mémoire.

Je remercie chaleureusement tous les relecteurs qui ont pris le temps de lire ce mémoire et de me faire part de leurs remarques constructives.

Pour finir, je remercie tout particulièrement les membres de ma famille, mes amis, mes camarades de master (M1 et M2), mes voisins du Rabot et Victor, pour m'avoir supportée et soutenue tout au long de cette période.

GLOSSAIRE

Écriture : Ressources graphiques qui permettent de représenter le langage et ses unités linguistiques (Fayol et Jaffré, 2008).

Graphème : La plus petite unité distinctive et/ou significative de la chaine écrite, composée d'une lettre, d'un groupe de lettres (digramme, trigramme), d'une lettre accentuée ou pourvue d'un signe auxiliaire, ayant une référence phonique et/ou sémique dans la chaine parlée. Il y a quatre graphèmes dans *chameau* : *ch*, *a*, *m*, *eau* ; dans *prends,* il y en a cinq : *p*, *r*, *en*, *d*, *s* (Catach, 1995 ; Ducard, Honvault et Jaffré, 1995).

Lemme : Forme arbitraire, mais conventionnelle, abstraite sur l'ensemble des formes d'un paradigme flexionnel (Polguère, 2003, cité dans Romary ; Salmon-Alt et Francopoulo, 2004). Nous prendrons comme lemme la forme à l'infinitif présent pour les verbes, le singulier pour les noms et le masculin singulier pour les adjectifs, pronoms, déterminants.

Orthographe : Convention graphique et sociale qui se sert d'une écriture pour donner à voir les signes linguistiques (Fayol et Jaffré, 2008).

Phonème : Plus petite unité distinctive de la chaine orale. Ensemble des sons reconnu par l'auditeur comme différent d'autres ensembles associés à d'autres ensembles. Il y a trois phonèmes dans *par* : /paR/ et deux phonèmes dans *homme* : /ɔm/ (Catach, 1995 ; Ducard *et al.*, 1995).

Phonographie : Ensemble des procédés qui permettent d'établir des correspondances entre des unités graphiques et les unités de l'oral (Fayol et Jaffré, 2008 ; Ducard *et al.*, 1995).

Sémiographie : Domaine de la linguistique de l'écrit qui regroupe l'ensemble des unités pourvues de sens (morphèmes et mots) (Ducard *et al.*, 1995).

CONVENTIONS D'ÉCRITURE

Pour la rédaction de ce mémoire, différentes conventions d'écritures sont utilisées. Celles-ci s'inspirent notamment des conventions adoptées par C. Blanche-Benveniste et A. Chervel (1978).

– Les productions ou parties de productions issues de notre corpus sont présentées entre guillemets et dans la police Agency FB. Les productions sont identifiées par un numéro attribué à l'élève scripteur, ce numéro est généralement inscrit en gras et entre parenthèses à la suite de la production. Pour une meilleure compréhension, les productions sont suivies d'une proposition de réécriture normée : « « […] le chat i no<letMF>u</letMF>le // <revision/>chat a le bébé chat […] » (**1363**, *le chat il miaule, chat a le bébé chat*).

– Les formes fléchies ou les formes produites en corpus sont inscrites en minuscules italiques : *chat*, lorsque la forme est jugée erronée, un astérisque est apposé devant la forme : **cha*. Les lemmes sont écrits en lettres capitales : CHAT, leur catégorie grammaticale peut éventuellement être précisée en indice : $CHAT_{NOM}$.

– Les lettres et graphèmes seront également en minuscules italiques : le graphème *a*.

– Les phonèmes, ainsi que les représentations phonologiques des mots sont présentés entre barres obliques au format API (Alphabet Phonétique International), les archiphonèmes se distingueront par le fait qu'ils seront présentés en majuscules : la représentation phonologique /ʃa/ et l'archiphonème /O/.

AVERTISSEMENT

Il est important de tenir compte du caractère évolutif du corpus à partir duquel s'effectue ce travail. La base de données étant complétée en parallèle, le nombre de productions a augmenté tout au long de l'étude, à l'instar des différents phénomènes à traiter. Les choix de transcription et de normalisation ont également pu évoluer entre les premières et les dernières analyses.

CHAPITRE 1 - INTRODUCTION

1.1. CONTEXTE GÉNÉRAL

Ce travail de mémoire a pour objet d'étude l'apport du traitement automatique de la langue (désormais TAL) à la constitution et à l'exploitation d'un corpus de productions d'élèves issues de classes de cours préparatoire (désormais CP) en France. Il est mené au sein du Laboratoire de Linguistique et Didactique des Langues Étrangères et Maternelles (Lidilem), rattaché à l'université Stendhal, Grenoble 3. Ce laboratoire présente des travaux relatifs aux disciplines des sciences du langage, principalement autour de la description linguistique, de la sociolinguistique, de l'acquisition, de la constitution et de l'exploitation de corpus, de la didactique des langues, du traitement automatique des langues, de l'étude des formes nouvelles d'interaction suscitées par les Technologies de l'Information et de la Communication. Ces travaux sont regroupés autour de trois axes : Descriptions linguistiques, TAL, corpus ; Sociolinguistique et acquisition du langage ; Didactique des langues, recherches en ingénierie éducative. Notre recherche sera menée au sein des axes 1 et 3.

Le corpus que nous étudierons a été collecté dans le cadre du projet national *Lire et écrire au CP*[1] (coordonné par Roland Goigoux et financé par la direction générale de l'enseignement scolaire (DGESCO), l'Institut français de l'Éducation (Ifé) et le laboratoire Acté (Clermont-Ferrand)), qui a pour objectif d'identifier les pratiques des enseignants de CP qui s'avèrent les plus efficaces au niveau de l'apprentissage de la lecture et de l'écriture, en particulier pour les élèves de milieu défavorisé.

Dans ce cadre, un groupe de chercheurs issu du groupe « écriture » du projet s'est constitué afin d'élaborer et d'annoter un corpus numérique de grande taille (au sens des corpus en contexte scolaire) de textes d'apprenants. Ce corpus est conçu pour être un corpus longitudinal permettant la description linguistique des productions écrites d'élèves en école primaire afin de rendre compte de l'évolution des procédés d'écriture entre le CP et le CM2. La constitution d'un corpus longitudinal est un travail qui se déroule sur plusieurs années, ce qui nécessite un suivi des classes et qui rend difficile la collecte des données. C'est pourquoi, sur les 131 classes de départ, seules les classes suivies par un membre du groupe « écriture »

[1] Nom complet du projet : *Lire et écrire au CP – Etude de l'influence des pratiques d'enseignement de la lecture et de l'écriture sur la qualité des premiers apprentissages.*

ou par un responsable du projet national ont été sélectionnées. Le corpus final concernera 60 classes, réparties en 5 académies : Grenoble (11 classes), Bordeaux (5 classes), Montpellier (8 classes), Lyon (19 classes) et Clermont-Ferrand (17 classes).

Dans le cadre du projet national, différentes épreuves sont proposées aux élèves à différents temps de leur scolarité. En septembre 2013 a eu lieu une première collecte. En juin 2014, une seconde collecte a permis de rassembler, entre autres, 2507 productions écrites relatives à une séquence de quatre images. Ces données constituent notre corpus.

1.2. LE RÔLE DU TAL

À terme, le corpus devrait contenir quelques milliers de productions. Un tel corpus ne peut pas être analysé manuellement. Le TAL a donc pour rôle d'aider linguistes, psycholinguistes et didacticiens à élaborer et à exploiter ce corpus, notamment en relevant les différents phénomènes d'études. À ce titre, nous nous plaçons résolument dans la même approche que celle développée par Kraif et Ponton (2007) à savoir une utilisation des technologies TAL les plus éprouvées dans un contexte relativement maitrisé pour une aide à l'analyse. Il ne s'agit donc en aucune manière de prétendre à une détection et à un diagnostic automatique complet des erreurs. Les études menées au sein du Lidilem par les porteurs du projet s'intéressent principalement à l'acquisition de l'écrit à travers l'observation des erreurs. C'est pourquoi, nous nous proposons d'élaborer un outil d'aide à la détection et à l'annotation automatique des erreurs au sens que nous venons d'évoquer.

Celui-ci s'inscrira dans un outil d'analyse plus global élaboré au sein du LIDILEM (figure 1). Cet outil permettra d'interroger notre corpus à l'aide d'outils de TAL. Il contiendra principalement une base de données agrémentée d'un module d'interrogation de cette base. La base de données contiendra en premier lieu les transcriptions annotées des productions. À l'exemple du corpus *Lancaster Corpus of Children's Project Writing* (LCCPW, Ivanic et McEnery 1996), chaque production sera accompagnée d'un scan de la production originelle de l'élève. Cette version sert notamment à renseigner l'utilisateur au niveau iconique comme la mise en page, la qualité de la calligraphie, etc., puisqu'une majeure partie de ces informations ne sont pas renseignées dans les transcriptions. Sera également ajoutée une proposition de réécriture normée permettant de donner des pistes de compréhension à l'utilisateur, ainsi que les métadonnées accompagnant chaque production et fournissant diverses informations comme l'âge de l'enfant, son sexe, la langue parlée à la maison, son milieu socio-culturel, etc.

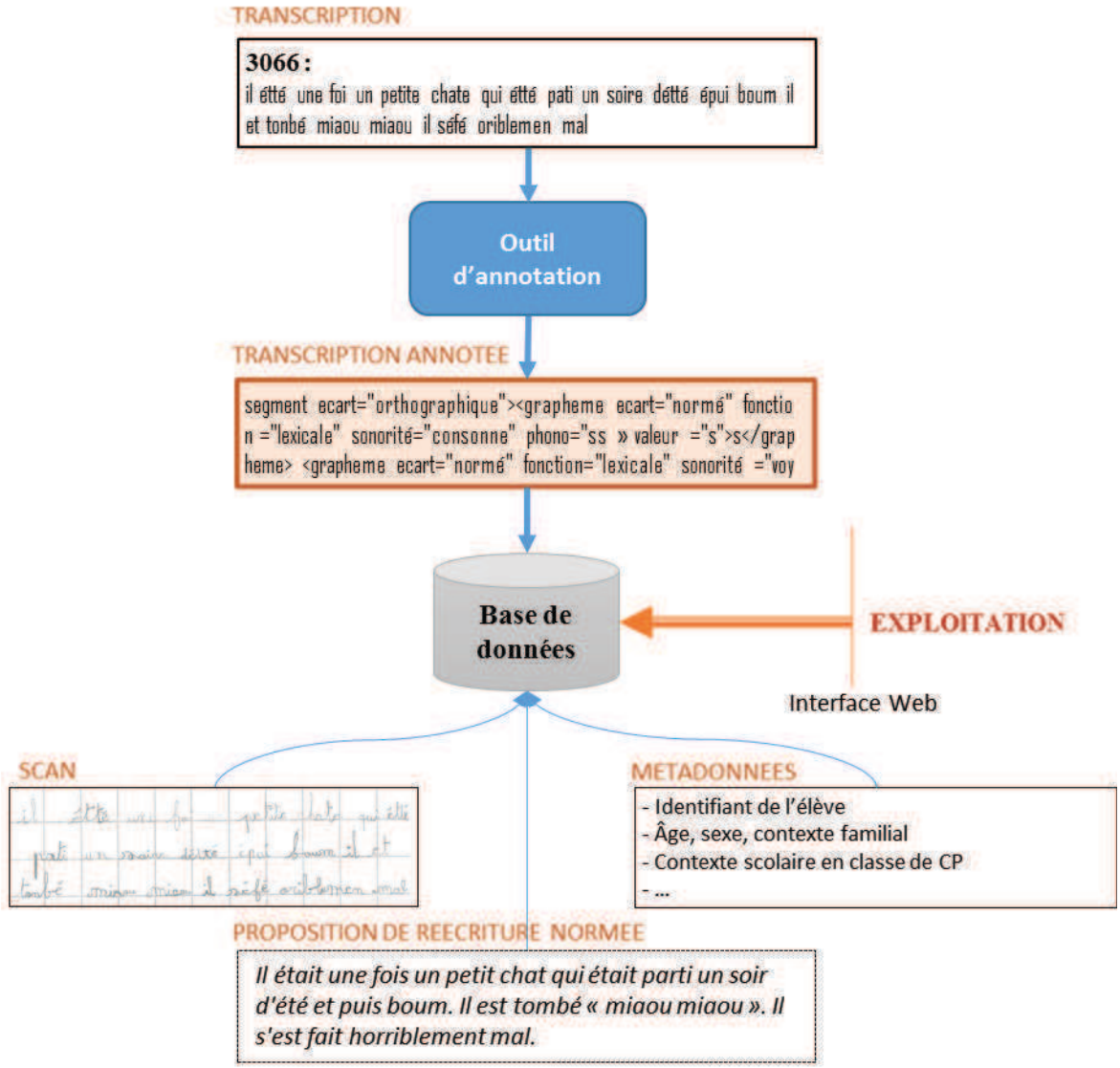


Figure 1. Schéma de la base de données contenant les productions

Une interface web utilisant des outils TAL permettra alors d'interroger la base de données ainsi constituée selon trois voies :

– La première possibilité est d'interroger la transcription annotée, une telle recherche permettra de relever les productions comportant un phénomène ou une erreur donnée. On pourra, par exemple, rechercher toutes les productions où la liaison a été marquée à l'initiale du mot, comme dans la production « il été une foi un chat qui ces promenet tout acou le chat tonbes parter il pleur le**s s**otre cerévei il lepren par le deau fin » (**563**, *Il était une fois un chat qui s'est promené. Tout à coup, le chat tombe par terre. Il pleure. Le**s a**utres se réveillent. Il le prend par le dos. Fin)*.

– Il est également possible de faire une recherche sur la forme normalisée, si l'on s'intéresse, par exemple, au nombre de fois que l'introducteur de récit *Il était une fois* a été utilisé. Cette version ayant été normalisée manuellement, toutes les occurrences de cette

formulation y seront répertoriées, même celles qui n'ont pas été reconnues dans la phase d'annotation, comme dans la production « **il étté une foi** un petite chate qui étté pati un soire détté épui boum il et tonbé miaou miaou il séfé oriblemen mal » (**3066**, ***Il était une fois*** *un petit chat qui était parti un soir d'été et puis boum. Il est tombé « miaou miaou ». Il s'est fait horriblement mal.)*.

– Enfin, il est également possible d'interroger ces deux versions de manière croisée. Dans le cas où l'on s'intéresse à la négation, par exemple, rechercher à partir des deux versions permet d'identifier les productions où la négation est correctement réalisée et celles où elle ne l'est pas comme dans : « une maman chat dormè et les chaton aussi ssofe un lui qui dormè pas tonba et pleurae la maman chat » (**1143**, *Une maman chat dormait et les chatons aussi sauf un. Celui qui* ***ne*** *dormait pas tomba et pleura. La maman chat<nonfini>*).

L'exploitation de la base de données ne concerne pas ce mémoire et ne sera pas plus détaillée ici.

1.3. CONCLUSION

Le travail de mémoire qui suit porte exclusivement sur la façon dont le TAL peut aider à l'élaboration et à l'exploitation d'un corpus scolaire, à travers la transcription des productions et leur annotation en terme d'erreurs. Ne pouvant pas tout traiter, les principales erreurs que nous traiterons sont les erreurs portées par les mots, c'est-à-dire des erreurs de niveau lexical. Il nous faudra, tout d'abord, réaliser une première observation de notre corpus, afin de mieux cerner les différents types d'erreurs que nous aurons à traiter. Puis, nous nous pencherons sur les différents systèmes de détection et de correction d'erreurs proposés en TAL et sur le type d'erreurs qu'ils permettent de corriger, avant de nous intéresser à la façon dont l'orthographe et l'acquisition de l'écrit sont décrits par les linguistes.

Nous proposerons alors une première version d'un schéma d'annotation d'erreurs qui puisse répondre au critère de calculabilité que nécessite le TAL tout en permettant une description linguistique des phénomènes rencontrés. Enfin, nous mettrons notre modèle en pratique afin d'annoter les segments présentant une phonologie identique au segment normé.

CHAPITRE 2 - CARACTÉRISTIQUES DU CORPUS

Afin de traiter au mieux notre corpus, il est nécessaire de le caractériser pour connaitre les phénomènes qui le composent. Nous nous pencherons tout d'abord sur la méthode de recueil du corpus et nous demanderons en quoi cette méthode peut influer sur son contenu. Par la suite, nous décrirons la nécessaire étape de transcription, qui a été la première étape de notre travail et qui permet de passer d'une production manuscrite à une production numérique. Enfin, une première série d'observations sera réalisée à partir d'un corpus restreint afin d'orienter la suite de notre recherche.

2.1. RECUEIL DU CORPUS

L'épreuve de production de texte réalisée en classe de CP est une épreuve collective, c'est-à-dire réalisée en classe entière. Lors de cette épreuve, quatre images (figure 2) étaient montrées aux élèves, qui disposaient ensuite de 15 minutes pour répondre à la consigne suivante : « *Aujourd'hui vous allez écrire chacun l'histoire d'un petit chat. Je vais vous montrer ce qui arrive à ce petit chat. Regardez bien les images. Vous allez écrire cette histoire ici. Si vous avez oublié l'histoire, vous pouvez retourner la feuille pour retrouver les dessins. Vous avez 15 minutes pour ce travail. Vous allez travailler seul ; personne ne vous aidera, par exemple à écrire un mot*. » (cf. annexe 2 pour la consigne complète). Lors de la rédaction, les élèves pouvaient, au besoin, consulter les images au tableau ou au dos de leur feuille.

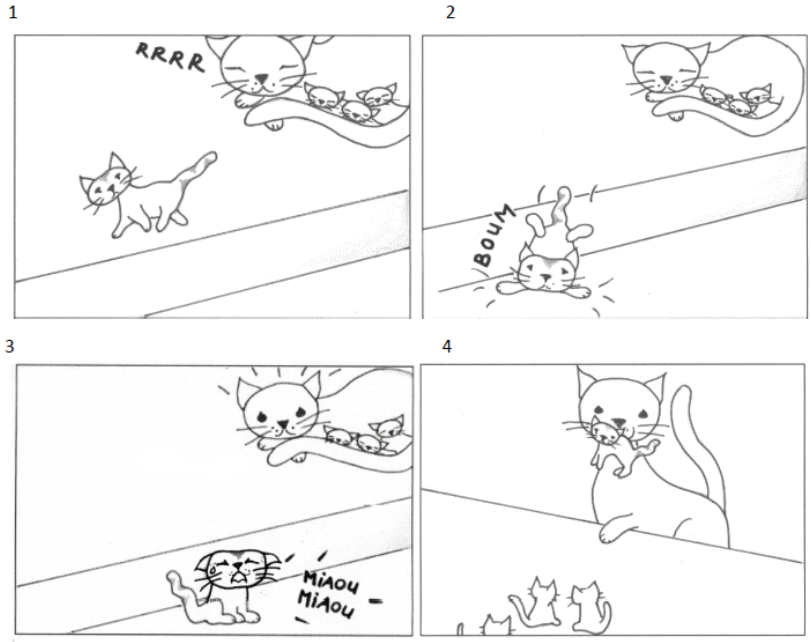


Figure 2. Images présentées aux élèves lors de leur production écrite

Les productions ont ensuite été centralisées au Lidilem où elles ont été scannées puis transcrites afin d'être enregistrées dans la base de données. À l'heure actuelle, seules 366 productions, soit 8 029 mots et 40 281 caractères, ont pu être rassemblées et transcrites. À terme, la base de données devrait en contenir 1170 pour l'année de CP. On observe sur cet ensemble de productions une moyenne de 21,9 mots par productions, soit une moyenne de 110,06 caractères par production. Cependant, ces chiffres peuvent être très variables d'une production à une autre. En effet, on observe un écart-type de 11 mots en moyenne. Alors que certains enfants n'ont rien produit (11 occurrences), d'autres ont écrit plus de 50 mots (7 productions), voire 60 mots (1 production).

2.2. DE L'ÉCRITURE MANUSCRITE À L'ÉCRITURE TAPUSCRITE : LA TRANSCRIPTION DU CORPUS

Manuel à l'origine, le corpus a d'abord été transcrit informatiquement afin d'en fournir une version numérisée. La version du corpus ainsi annotée permettra alors d'analyser linguistiquement les productions afin d'étudier les processus d'écriture et d'en mesurer l'évolution, dans le cadre du corpus longitudinal.

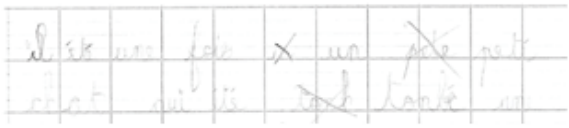
Cependant, comme le précise M.-L. Elalouf (2005), cette étape implique souvent une interprétation du corpus, tant dans l'identification des lettres que des mots. Transcrire et annoter un corpus consiste donc à appliquer les meilleurs choix interprétatifs. La transcription du corpus implique également de faire des choix quant aux informations à numériser (textes, mises en pages, ratures…). L'analyse de ces transcriptions se faisant par des techniques automatiques, il est primordial que cette transcription soit rigoureuse et homogène sur l'ensemble du corpus. En vue de cette homogénéisation, un guide a été élaboré expliquant les choix réalisés (annexe 3).

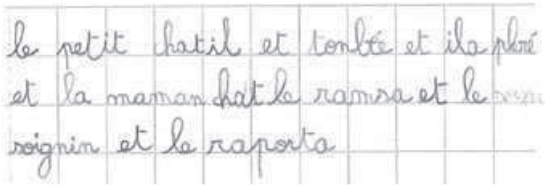
Les marques de révisions

La transcription du corpus a été pensée pour une utilisation par des didacticiens et non des généticiens du texte. Cette perspective a influencé le choix des phénomènes à annoter. C'est pourquoi les ratures, réécritures et traces de gomme sont relevées mais ne sont pas différenciées. Elles seront toutes considérées comme des marques de révision de la part de l'enfant et seront relevées à l'aide de la balise <revision/>. En effet, d'un point de vue purement linguistique, centré sur la production finale, il ne nous intéresse pas de les distinguer.

« […] le chat i no<letMF>u</letMF>le // <revision/>chat a le bébé chat a […] » (**1363**, *Le chat il miaule. Chat a le bébé chat a*)

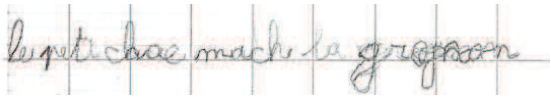
« il été une fois <revision/> un <revision/> peti / chat qui été <revision/> tonbé une […] » (**1560**, *Il était une fois un petit chat qui était tombé d'une*)





« le petit chat il et tonbée et il a pleré / et la maman <revision/>chat le ramsa et le <revision/>soignin et le raporta » (**2944**, *Le petit chat, il est tombé et il a pleuré et la maman chat le ramassa et le soigna et le rapporta.*)

« le peti chae mach la <revision/><illisible/ » (**2533**, *Le petit chat marche. La*)



Dans le cas où une analyse plus fine de ces traces serait nécessaire, il faudra modifier la balise que nous utilisons.

Les retours à la ligne

Contrairement à ce que propose M.-L. Elalouf (2005), la sortie de la transcription ne respecte que peu l'iconicité du texte, une lecture facilitant un traitement automatique y étant privilégiée. Toutefois, le scan de la version manuscrite permet de consulter la forme visuelle du texte.

Cependant, il nous a paru important de distinguer les retours à la ligne perçus comme volontaires, et qui peuvent donc être porteurs de sens, des retours à la ligne contraints spatialement par l'extrémité de la feuille. Les premiers seront notés par un slash double //, tandis que les deuxièmes le seront par un slash simple /.

« le chat elle marche. // le chat étonbé. // le chat elle miole. // la maman il la sové le chat peti. » (**1361**, *Le chat, il marche. Le chat est tombé. Le chat, il miaule. La maman, elle a sauvé le petit chat.*)

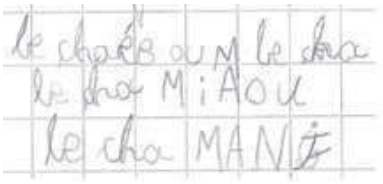
« Le petit chat et tonder il c'est / fait male au pate et a la téte il a / fait maou sa maman vien le / récupére le petit chat. » (**1354**, *Le petit chat est tombé. Il s'est fait mal aux pattes et à la tête. Il a fait miaou. Sa maman vient le récupérer, le petit chat.*)

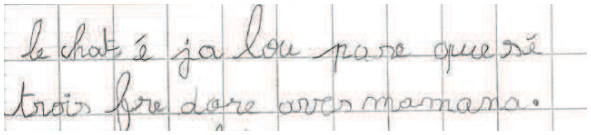
Anomalies graphiques

Il nous a également paru important de relever les anomalies graphiques et les hésitations, erreurs qui ne seront plus prises en compte par la suite. Les lettres mal formées sont relevées par une balise spécifique, ainsi que les lettres difficiles d'interprétation. Dans ces derniers cas, toutes les interprétations plausibles sont marquées, en mettant en avant la plus probable. Rappelons tout de même que, malgré la présence de ces balises qui marquent une certaine distance par rapport à la lecture que nous faisons de notre corpus, transcrire implique toujours une certaine part d'interprétation.

« le cha<letMF>t</letMF> dor et le bébé cha<letMF>t</letMF> miol / et le cha<letMF>t</letMF> il <letMF>t</letMF>onbe Boum [...] » (**1297**, *Le chat dort et le bébé chat miaule et le chat il tombe. Boum*)

« le ch<letMF>a</letMF>deur // le cha<revision/>seréfêille // le chaée OUB le cha // le cha MIAOU // le cha MAN<J|G> » (**2930**, *Le chat et BOUM le chat. Le chat MIAOU. Le chat mange.*)



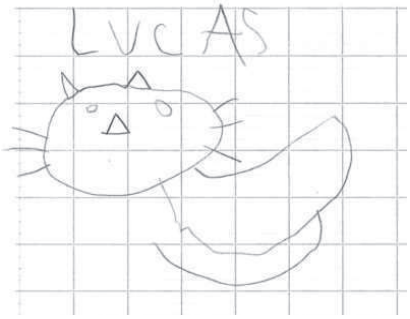


« le chat é ja lou pa<s|r>e que sé / trois fre dore aves mamana. » (**1128**, *Le chat est jaloux parce que ses trois frères dorment avec maman.*)

Segments non transcrits

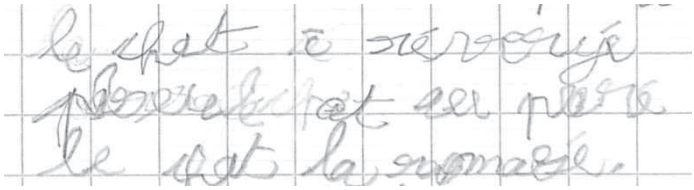
Enfin, les segments que nous n'étions pas en mesure de transcrire ont également fait l'objet de balises spécifiques, à savoir les dessins et les segments composés de lettres non identifiables. Nous avons également fait le choix de ne pas transcrire les prénoms des élèves visibles dans les productions autrement que par une balise <prenom/>. De même, les suites de

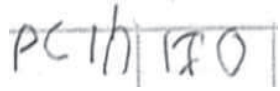
lettres détachées non analysables automatiquement, ont été transcrites par la balise <letDET/>.



« <prenom/> // <dessin/> » (**1288**)

« [...] le chat <letMF>ê</letMF> réveryé / <illisible/> <revision/> at <illisible/> pl<letMF>e</letMF>ré / le chat la ramasé. [...] » (**1341**, *Le chat est réveillé, a pleuré. Le chat l'a ramassé.*)





« <letDET/> » (**1553**)

Ces balises servent principalement à marquer des segments que nous ne traiterons pas, seules les marques de retour à la ligne seront utilisées dans certaines conditions. C'est pourquoi, dans la suite de notre travail, nous nous baserons généralement sur une version des productions exemptées de ces balises.

2.3. BRÈVE ÉTUDE LEXICOMÉTRIQUE

Afin de clarifier l'analyse et le travail qui vont suivre, nous appellerons **segment** toute séquence de lettres séparée par des frontières de mots comme le blanc graphique, le retour à la ligne ou la ponctuation. Ce terme pourra correspondre ou non à un mot bien formé en langue française et sera utilisé dans le cadre des productions des élèves. Le terme **forme** désignera toute séquence de lettres séparées par des frontières de mots telles qu'on en trouve dans les lexiques de formes fléchies. Cette notion mériterait d'être plus approfondie pour un travail ultérieur, mais nous parait suffisante dans le cadre de ce mémoire portant principalement sur l'orthographe et la segmentation.

La plupart des observations réalisées sur notre corpus étant manuelle, nous avons dû élaborer des sous-corpus de taille restreinte. Un premier sous-corpus a été élaboré en sélectionnant, à l'aide d'un script, l'ensemble des productions finissant par le chiffre 6, méthode qui devait permettre d'obtenir des productions de chaque classe et donc issues d'enseignements différents pour rendre compte de la variété qui peut exister dans le corpus. Ce corpus contenait au départ 17 productions (cf. 2.4.), soit 393 segments. Puis, en raison de

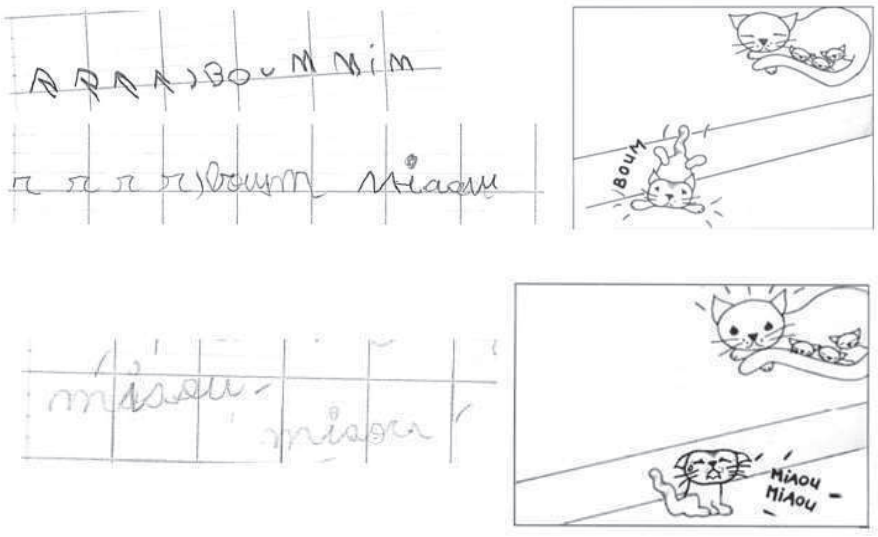
l'entrée de nouvelles transcriptions dans la base de données, il en contint 20 lors de sa deuxième utilisation (cf. chapitre 8.), soit 471 segments. Cependant, certaines analyses, comme la lexicométrie, tout en étant manuelles, nécessitent un corpus un peu plus conséquent. C'est pourquoi, un deuxième sous-corpus a été élaboré contenant 40 productions prises au hasard dans notre corpus.

La méthode de recueil utilisée a conduit un certain nombre d'élèves à produire des textes qui s'apparentent à une description d'images. Ceci nous amène à émettre l'hypothèse que le vocabulaire employé par ces élèves se limite souvent à des mots permettant de désigner les personnages et leurs actions. L'ensemble des productions devrait donc présenter un vocabulaire récurrent et un usage courant de certains noms comme CHAT et CHATON ou encore de verbes d'action comme TOMBER, REVEILLER ou PLEURER.

Afin de vérifier cette hypothèse, nous avons utilisé le deuxième sous-corpus contenant 40 productions, soit 942 formes, pour lequel nous avons proposé une correction afin de nous concentrer sur les caractéristiques lexicales de notre corpus et non sur les erreurs. Cette correction est envoyée à TreeTagger (Schmid, 1994), outil qui permet d'analyser la catégorie et le lemme d'une forme et qui présente l'avantage d'être libre de droit. Puis, deux scripts écrits en Perl[2] permettent de compter le nombre d'occurrences de chaque forme et de chaque lemme et de les classer dans l'ordre décroissant de fréquence. On obtient ainsi deux tableaux, l'un contenant les formes et l'autre les lemmes (annexe 4).

À partir de ces tableaux, on constate que le lemme CHAT a été utilisé 77 fois sur un total de 942 formes, soit une fréquence de 8,2%, ce qui signifie qu'il est présent à hauteur de 1 lemme sur 12. Les lemmes TOMBER, PLEURER, REVEILLER et CHATON se retrouvent respectivement en 9^{e}, 12^{e}, 14^{e} et 26^{e} position et apparaissent à hauteur de 29, 17, 22 et 7 occurrences, ce qui tend à montrer une certaine récurrence du vocabulaire. On note également la réutilisation fréquente des onomatopées BOUM, MIAOU et RRRR présentes sur les images, MIAOU arrivant, en effet, en 13^{e} position. Certains élèves vont jusqu'à copier certains éléments du dessin, extérieurs à la graphie de l'onomatopée, comme le trait précédant BOUM ou ceux autour de MIAOU :

[2] Langage de programmation. https://www.perl.org/



Nous nous sommes également demandé si le vocabulaire utilisé dans notre corpus était varié ou « riche » (Lebart et Salem, 1994). Pour ce faire, nous nous sommes basé sur la proposition de correction de notre sous-corpus de 40 productions. Nous n'avons donc pas prétention à l'exactitude, la correction étant elle-même sujette à interprétation. De plus, au vu du nombre restreint de formes contenues dans notre corpus, notre but est uniquement de dégager quelques tendances qui nous donneront éventuellement quelques pistes de traitement, mais qui nécessiteront d'être vérifiées sur un corpus plus important.

Cette mesure ne peut être calculée que s'il y a comparaison de notre corpus avec d'autres corpus de référence. Nous en avons choisi trois. Notre corpus est un corpus narratif d'enfants qui nous semble encore très proche de l'oral. Nous avons donc choisi comme premier corpus de référence une transcription écrite d'un récit oral fait par une enfant de 5 ans. Ce corpus est tiré du projet « Traitement de Corpus Oraux en Français » (TCOF) (Benzitoun, Fort et Sagot, 2012). Dans ce même corpus, nous avons également sélectionné une transcription d'une production orale d'adulte. Ce corpus ne comportant aucun récit produit par un adulte, nous avons choisi une partie de cours universitaire portant sur les comptines. Enfin, nous avons sélectionné un texte narratif écrit d'adulte racontant l'histoire d'un chat, intitulée « Le petit chat désobéissant » et adaptée d'un conte russe. Les deux derniers textes nous semblent refléter le vocabulaire d'un adulte, tandis que le premier nous permettra une comparaison avec d'autres productions enfantines. Pour que les mesures effectuées sur ces différents corpus soient comparables, leur longueur doit l'être également. Notre sous-corpus contenant 942 formes et le corpus oral d'enfant en contenant 894, nous limiterons les deux autres corpus, plus longs, aux 900 premières formes.

En réutilisant les scripts précédents, nous avons pu extraire un tableau de fréquence de formes et un tableau de fréquence de lemmes pour chaque corpus.

Corpus corrigé		Corpus oral enfant		corpus oral adulte		Corpus écrit adulte	
chat	74	elle	55	de	32	le	27
il	67	il	49	les	31	de	23
le	65	est	45	et	25	il	19
et	55	et	44	la	19	la	17
maman	44	là	43	ils	16	que	17
sa	31	c'	28	qui	16	un	16
se	26	le	26	on	15	tu	13
petit	23	la	20	que	15	petit	13
miaou	22	Belle	19	dans	15	est	13
est	20	en	17	le	14	dit	13
Le	19	euh	16	des	14	se	13
la	19	a	12	d'	14	et	12
un	18	se	12	pas	13	je	12
tombe	16	voit	12	tout	13	les	11
pleure	13	Bête	12	ça	13	s'	11
a	12	de	11	est	13	pas	10
fait	12	une	11	enfants	11	une	10
de	10	était	10	qu'	11	chaton	9
bébé	10	son	9	voilà	11	lapin	9
boum	10	avec	9	c'	11	écureuil	8

Tableau 1. Extrait du tableau des formes associées de leur fréquence

La première remarque que l'on peut faire devant ce tableau consiste à relever la prédominance des formes lexicales dans les formes les plus fréquentes de notre corpus par rapport aux autres corpus, dont les formes les plus fréquentes sont généralement des formes grammaticales. Plusieurs raisons peuvent être évoquées pour l'expliquer comme la redondance des formes lexicales de notre corpus ou encore une faiblesse grammaticale dans ce même corpus. Toutefois, comprendre la cause de ce phénomène n'est pas l'objet de notre étude.

Un des indices fréquemment utilisés pour mesurer la richesse lexicale d'un texte consiste à faire le ratio entre le nombre de mots total d'un texte, autrement appelé occurrences, et le nombre de mots différents, appelé formes. Cette mesure est appelée *rapport formes / occurrences* (Véronis, 2008) ou encore *type / token ratio*. Nous avons évalué ce rapport à la fois pour les lemmes et pour les formes de chacun de nos corpus :

	Corpus corrigé	Corpus oral enfant	corpus oral adulte	Corpus écrit adulte
Rapport formes/occurrences	0,21	0,25	0,33	0,41
Rapport lemmes/occurrences	0,15	0,19	0,28	0,28

Tableau 2. Mesure des rapports formes / occurrences et lemmes/occurrences

Selon cette mesure, notre corpus semble moins riche lexicalement parlant que les corpus d'adultes. Toutefois, une autre mesure envisagée de la « richesse lexicale » repose sur

le taux d'hapax (forme présente de façon unique dans un corpus) comparé au nombre de lemmes. Pour chaque corpus, ce taux est reporté dans le tableau 3.

	Corpus corrigé	**Corpus oral enfant**	**corpus oral adulte**	**Corpus écrit adulte**
Taux d'hapax/lemmes	0,66	0,69	0,9	0,67

Tableau 3. Mesure de la richesse lexicale à partir des hapax

À partir de cette mesure, seul le texte narratif adulte semble significativement plus riche, nous ne pouvons donc conclure à une pauvreté lexicale ou à une absence de variabilité dans notre corpus. Nous pouvons néanmoins retenir l'importante fréquence de certaines unités lexicales. Ces éléments nous semblent constitutifs de notre corpus et pourront être réutilisés dans la suite de notre travail.

2.4. OBSERVATION DES ERREURS

Une première observation des erreurs de notre corpus devrait nous permettre d'extraire les caractéristiques les plus saillantes sur lesquelles reposera la suite de notre travail. Ces premières observations étant manuelles, elles reposent sur le sous-corpus des productions finissant par 6, composé de 17 productions.

2.4.1. LES ERREURS INTRA-MOTS

Ce sous-corpus va nous permettre une première observation globale à partir des opérations de bases sur une chaine de caractères, à savoir l'omission, l'insertion et la substitution, à l'exemple des premiers correcteurs orthographiques pour scripteurs experts (cf. chapitre 2).

Nous avons commencé l'étude de ces erreurs en prenant pour unité la lettre. Cependant, il est rapidement apparu que prendre cette unité ne permettait pas de traiter aisément les erreurs incluant plusieurs de ces unités, comme le cas où la suite de lettre *ai* est remplacée par la lettre *é*, dans « eté » (**1156**, *était*). L'unité lettre ne nous permet donc pas de prendre en compte l'implication de la phonologie dans notre système d'écriture. Nous avons alors élargi cette unité aux lettres et groupes de lettres, qui peuvent comprendre deux, voire trois lettres. De plus, certaines erreurs ne sont également portées que par le signe diacritique, généralement l'accent, comme dans « lève » (**2986**, lève) tandis que d'autres portent sur des syllabes entières, à l'exemple de « en dus » (**1226**, *entendu*). C'est pourquoi, pour cette première étude du moins, nous avons volontairement gardé les unités de mesure les plus

larges possibles, en ne conservant pour tout unité que les lettres et groupes de lettres et les diacritiques.

Les observations faites ont été reportées dans des tableaux dont le détail complet est disponible en annexe (annexe 5), mais nous pouvons en rapporter ici une synthèse. Les 17 productions étudiées comptent 393 segments dont 147 comportent une ou plusieurs erreurs. Parmi ces erreurs, on dénombre : 100 suppressions, 28 insertions, 95 substitutions et une inversion. Pour chacune d'elle, on précisera si elle porte sur une lettre ou un groupe de lettres (« son_ », **1986**, *son**t***) ou un accent (diacritique) (« ét », **1156**, ***é**t*).

Cependant l'unité lettre ou groupe de lettres et très englobante et il nous a paru important de faire quelques distinctions. On distinguera ainsi les erreurs qui portent sur les consonnes doubles (« diféran », **1156**, *di**ff**érent*) des autres. Pour ces autres erreurs, une différence est faite selon leur implication dans la valeur phonique du mot. On distingue ainsi les lettres prononcées (« mongée », **2986**, *manger*) des lettres muettes (« dore », **1346**, *dor**t***). Enfin, pour chaque type d'erreur est indiqué s'il altère la valeur phonique (Oral altéré) ou non (Oral conservé). L'erreur « tondé », prononcé /tɔ̃de/ (**1336**, *tom**b**é* prononcé /tɔ̃be/) modifie la valeur phonique, tandis que l'erreur « tonbé » (**1556**, *t**om**bé*) n'entraine aucune modification au niveau de la prononciation.

Unité de l'erreur	**Omission**		***Oral altéré***	***Oral conservé***	***Total***
	Lettre ou groupe de lettres	Prononcé	21		21
		Muet		64	64
		Fin de mot		61	
		Milieu ou début de mot		3	
		Consonne double (omission d'une lettre)		6	6
	Diacritique	Accent	5	4	9
	Total		**26**	**74**	**100**

Tableau 4. Phénomènes d'omission observés dans le sous-corpus

En plus des distinctions exposées plus haut, nous avons également reporté le nombre d'omissions affectant les lettres muettes en milieu ou en début de mot (« oriblemen », **3066**, ***h**orriblement*) et le nombre des celles affectant les lettres muettes en fin de mot (« oriblemen_ », **3066**, *horriblemen**t***). Ces erreurs sont beaucoup plus nombreuses.

Unité de l'erreur	**Insertion**		***Oral altéré***	***Oral conservé***	***Total***
	Lettre ou groupe de lettres	Prononcé	11		11
		Liaison	1		
		Muet		6	6
		Consonne double (insertion d'une lettre)	1	4	5
	Diacritique	Accent	4	3	7
	Total		**15**	**13**	**28**

Tableau 5. Phénomènes d'insertion observés dans le sous-corpus

Au niveau des phénomènes d'insertion, il nous a semblé intéressant de relever la transcription de la consonne de liaison, audible à l'oral (« des **s**otre », **1156**, *des autres*). Ce phénomène n'apparait qu'une fois dans notre sous-corpus, mais apparait plusieurs fois dans le corpus.

Unité de l'erreur	**Substitution**		***Oral altéré***	***Oral conservé***	***Total***
	Lettre ou groupe de lettres	Prononcé	11	76	87
		Phonologie identique		76	
		Phonologie proche	6		
		Graphie proche	3		
		Autre	2		
		Muet		5	5
	Diacritique	Accent	3		3
	Total		**14**	**81**	**95**

Tableau 6. Phénomènes de substitution observés dans le sous-corpus

Au vu du nombre conséquent de substitutions apparaissant dans notre sous-corpus, il nous a paru pertinent de préciser quelque peu la catégorie des substitutions affectant les lettres ou groupes de lettres prononcés. Nous avons relevé des substitutions n'affectant pas la phonologie (Phonologie identique), comme « t**on**bé » (**1556**, *t**om**bé*) et des substitutions affectant la phonologie. Parmi ces dernières, certaines portent plutôt sur une relation de proximité phonique (« ton**d**é », **1336**, *tom**b**é*), d'autres sur une relation de proximité graphique (« m**on**gée », **2986**, *manger*), bien que ce dernier critère soit discutable.

Enfin, il existe un dernier type d'erreur pour lequel un seul cas a été relevé, tout au moins dans le sous-corpus étudié. Il s'agit de l'inversion de lettres, relevée dans le segment « ap**ér** » (**3006**, *ap**rè**s*).

Ces tableaux sont le reflet d'une première réflexion empirique et certains phénomènes ont, suite à des discussions avec le groupe du projet, bénéficié d'un classement différent. Par exemple, l'absence d'une consonne lorsqu'une consonne double est attendue peut être vue comme une suppression d'une lettre dans un digramme (séquence de deux lettres), comme

une substitution d'un graphème à un autre, ou encore comme une suppression de graphème si l'on considère les deux consonnes comme deux graphèmes différents.

Il est rapidement apparu que les erreurs les plus fréquentes sont des erreurs de substitution et de suppression. Plus particulièrement, il s'agit d'erreurs de substitution de graphèmes à des graphèmes de même valeur phonique, à raison de 81 substitutions sur les 95 de notre échantillon, et de suppressions de lettres muettes, à raison de 64 suppressions sur les 100. Ces deux modifications n'affectent pas la phonologie. Il nous est donc paru pertinent de distinguer les erreurs affectant la forme phonologique des erreurs n'ayant pas d'incidence sur cette forme (tableau 7).

Type d'erreurs	**Oral altéré**	**Oral conservé**	**Total**
Omission	26	74	100
Insertion	16	13	29
Substitution	14	81	95
Inversion	1		1
Total	**57**	**168**	**225**

Tableau 7. Synthèse des erreurs selon leur implication dans la chaine orale

En outre, à première vue, les phénomènes de flexion et d'accord qui ne se perçoivent pas dans la chaine sonore semblent très peu présents dans ce corpus, ce qui nous amène à penser qu'à ce stade, les enfants ont pour principal souci de transcrire fidèlement la chaine sonore. Notre analyse sera donc principalement portée sur la distinction entre chaine sonore fidèlement transcrite et chaine sonore partiellement transcrite.

Enfin, notons également que la correction nécessitée par l'étape précédente a rendu saillants quelques phénomènes d'erreurs fréquents. En effet, on a pu constater que la forme *chat* au singulier était régulièrement bien orthographiée, à hauteur de 63 sur 74 occurrences, alors que la forme *tombe* était le plus souvent mal orthographiée. Celle-ci est orthographiée *tombe* dans seulement 3 cas sur 16, alors qu'elle est orthographiée **tonbe* 8 fois. Il semble donc qu'il y ait des formes susceptibles de présenter plus d'erreurs que d'autres.

2.4.2. LES UNITÉS DE L'ÉCRIT

Outre les erreurs portant sur l'orthographe des mots, notre corpus présente également pour particularité une segmentation en unité de l'écrit pas toujours normée. Or, comme nous le verrons par la suite, il est important pour qu'un système de TAL soit performant que ces unités soient bien distinguées. Nous nous intéresserons donc à la façon dont les jeunes scripteurs ont segmenté leurs productions en unités lexicales et en unités syntaxiques. Pour

clarifier notre propos, nous emploierons les termes *mots* et *phrases*, laissant de côté les problèmes que posent ces deux notions.

2.4.2.1. SEGMENTATION EN MOTS

La segmentation en mots a été observée à l'aide du même sous-corpus de 17 productions que précédemment. Deux types d'erreurs de segmentation sont étudiés :

– la **sur-segmentation** ou hypersegmentation : lorsqu'un mot est divisé en plusieurs segments différents ;

– l'**agglutination** ou hyposegmentation : lorsque plusieurs mots sont regroupés en un seul segment.

Les résultats de cette observation sont regroupés dans le tableau 8.

Segmentation	
Hypersegmentation	6
Hyposegmentation	19
Omission de l'élision	7
Total	25

Tableau 8. Erreurs de segmentation dans le sous-corpus

Ces phénomènes ne sont pas très nombreux mais sont primordiaux car ils risquent de corrompre l'analyse des erreurs lexicales.

2.4.2.2. SEGMENTATION EN PHRASES

Les analyses syntaxiques permettent bien souvent de désambigüiser les unités lexicales ou leur catégorie. C'est pourquoi il est important de pouvoir repérer les phrases dans les productions. Nous ne nous sommes pas intéressée ici à la composition des phrases mais plutôt aux marqueurs permettant de les identifier. Il apparait rapidement que bien que n'utilisant pas encore tous les marqueurs d'un scripteur adulte, un certain nombre de jeunes scripteurs emploient déjà certaines stratégies pour segmenter leur texte.

La ponctuation

Observons tout d'abord l'usage du point et de la ponctuation, principal marqueur d'une fin de phrase dans les productions de scripteurs experts. Certains l'utilisent effectivement pour segmenter leurs phrases, à l'instar de la production :

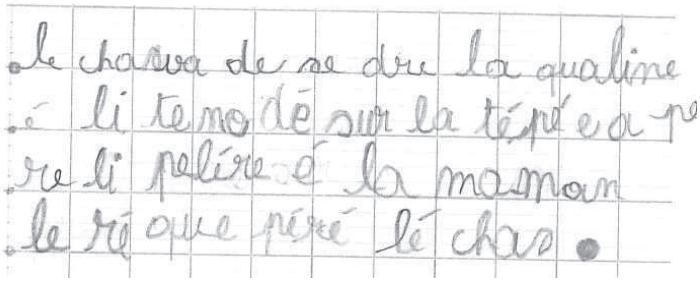
> « Le petit chat et partie de son lit. Mai boum il tombe et il pleurer. Sa maman et sai fraire se raivaya et il le voillér pleurer. Et miantenan sai le matin. La mamans les porte pour les sortir du lit. Fin » (**586**, *Le petit chat est parti de son lit. Mais boum, il tombe et il*

pleurait. Sa maman et ses frères se réveillèrent et ils le voyaient pleurer. Et maintenant c'est le matin. La maman les porte pour les sortir du lit. Fin)

Mais beaucoup l'emploient comme un marqueur de fin de texte, comme dans la production :

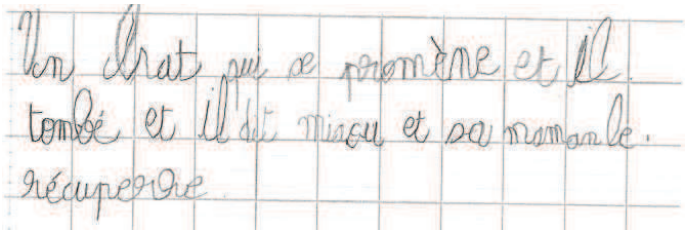
« Le petit cha se sove en couran il tombe il se fait male sa maman se rèvie elle prena le petit cha il ètai sovèe. » (**584**, *Le petit chat se sauve en courant. Il tombe. Il se fait mal. Sa maman se réveille. Elle prit le petit chat. Il était sauvé.*)

Le point peut également servir à marquer un début ou fin de ligne, sans rapport avec la phrase.



« le chasva de se dr<e|i> la qualine / . é li teno dé sur la tépé e a pe / . re li pelire é la maman / . le ré que péré lé chas. » (**3071**, *Le chat s'en va <incomprehensible/>. Et il tombe sur la tête et après il pleure et la maman le récupère le chat.*)

« un chat qui ce promène et il. / tombé et il dit miaou et sa maman le./ récuperre. » (**657**, *Un chat qui se promène et il tombe et il dit miaou et sa maman le récupère.)*



Les majuscules

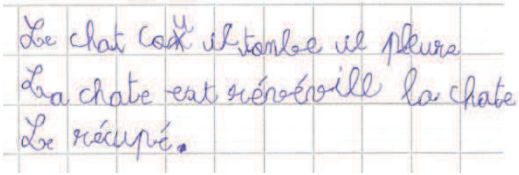
Si la ponctuation n'est pas un indicateur fiable pour repérer les phrases dans les productions de scripteurs débutants, les majuscules n'en sont pas non plus. En effet, dans notre corpus, peu de majuscules sont présentes, à l'exemple de la production **1156**, qui n'en contient aucune :

« il eté diféran des sotre à prè / il tombe ēt aprè sa fait / boum et aprè il fait miaou deux / foi et sa mère la trape. » (**1156**, *Il était différent des autres. Après il tombe et après ça fait boum et après il fait « miaou » deux fois et sa mère l'attrape.*)

Lorsqu'elles le sont, elles ne sont pas toujours utilisées à bon escient. Dans la plupart des cas, l'on trouve une majuscule unique à l'initiale du texte, comme dans l'exemple :

« Le jéan chat dore avec c'est / chaton le petite chat marche sur / la route et tonbe sur la route / et pui il révèille le jean chat / et pui le petite <revision/>chat cris miaou</revision> miaou / et pui le jean chat » (**1350**, *Le géant chat dort avec ses chatons. Le petit chat marche sur la route et tombe sur la route et puis il réveille le géant chat et puis le petit chat crie miaou miaou et puis le géant chat <nonfini>*)

Mais l'on retrouve également des textes qui, comme pour la ponctuation, comportent des majuscules à chaque début de ligne, bien que ce puisse être le milieu de la phrase :



« Le chat cou il tonbe il pleure / La chate est révévill la chate / Le récupé. » (**3033**, *Le chat court. Il tombe. Il pleure. La chatte est réveillée. La chatte le récupère.*)

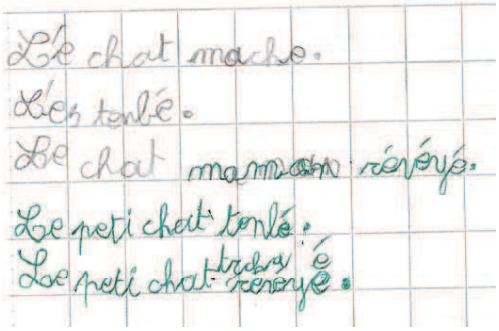
Enfin, l'on trouve quelques textes où l'utilisation des majuscules semble maitrisée.

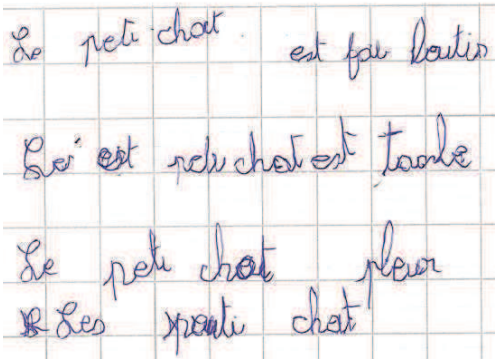
« Le petit chat quite sa maman et c'est / frère. Mais... Boume il tombe. Il pleur / miou miaou. Ouf sa maman revien le / cherché est elle le ramène. » (**3007**, *Le petit chat quitte sa maman et ses frères. Mais... Boum il tombe. Il pleure miaou miaou. Ouf sa maman revient le chercher et elle le ramène.*)

Le retour à la ligne

Notons également qu'une stratégie couramment utilisée dans les productions de scripteurs débutants est le retour à la ligne pour chaque phrase et éventuellement l'utilisation d'une majuscule en début de ligne et d'un point en fin de phrase :

« Le chat mache. // Les tonbé. // Le chat maman révéyé. // Le peti chat tonbé. // Le peti chat trcha é révéyé. » (**3037**, *Le chat marche. Il est tombé. Le chat maman est réveillé. Le petit chat est tombé. Le petit chat <incomprehensible> est réveillé.*)





« Le peti chat est fab boutis // Le'est p<letMF>e</letMF><letMF>t</letMF><letMF>i</letMF> chat est tanbe // Le peti chat pleur // <illisible/>Les pe<revision/>ti chat » (**3036**, *Le petit chat est <incomprehensible>. Le petit chat est tombé. Le petit chat pleure. Les petits chats<nonfini>*)

Les connecteurs

Enfin, on trouve de nombreux connecteurs, principalement *et*, placés en position de séparateurs de phrases, remplaçant la ponctuation :

> « le chat quoure é il tonbe // é il plére é sa maman / lui faiun qualin é le chat / na plumale » (**573**, *Le chat court et il tombe et il pleure et sa maman lui fait un câlin et le chat n'a plus mal.*)

> « il eté diféran des sotre à prè / il tombe <letMF>é</letMF>t aprè sa fait / boum et aprè il fait miaou deux / foi et sa mère la trape. » (**1156**, *Il était différent des autres. Après il tombe et après ça fait boum et après il fait « miaou » deux fois et sa mère l'attrape.*).

Ces exemples tendent à montrer que, bien que n'utilisant pas les marqueurs de phrases classiquement utilisés par des scripteurs experts, certains jeunes scripteurs opèrent déjà une segmentation en unités syntaxiques. De plus, les erreurs présentes au niveau de la segmentation en phrases sont peu nombreuses. Les problèmes liés au repérage des unités syntaxiques relèvent plutôt d'absence de marqueur que d'erreurs sur ces marqueurs.

Tous les phénomènes et erreurs repérés dans cette partie, sont autant de pistes pour la suite de notre travail. Maintenant que nous avons une vision plus claire des phénomènes que nous aurons à traiter, nous pouvons nous pencher sur les méthodes automatiques existantes qui nous permettront de les résoudre. Plus spécifiquement, il nous faut trouver des méthodes permettant de repérer les phrases de notre corpus, de détecter et de typer les erreurs d'orthographe lexicale et celles de segmentation.

CHAPITRE 3 - ÉTAT DE L'ART

De nombreuses applications en TAL nécessitent un prétraitement dont l'objectif est de normaliser le corpus en présence en gérant les cas particuliers pour faciliter les traitements ultérieurs. Cette étape de normalisation permet de repérer les unités lexicales et syntaxiques afin de les normer selon les besoins de l'application. Cette étape inclut souvent des correcteurs et des outils de segmentation. Les questions d'orthographe et de segmentation sont donc des problèmes bien connus en TAL pour lesquels de nombreuses méthodes de résolution ont été avancées.

Après une rapide présentation du domaine du TAL, nous nous pencherons sur les méthodes envisagées pour résoudre les problèmes de segmentation tant syntaxiques que lexicaux. Nous envisagerons ensuite un panorama des techniques de corrections lexicales les plus utilisées ou les plus adaptées à notre corpus. Si peu de travaux ont été réalisés dans le domaine du TAL sur des corpus d'acquisition de langue maternelle, des travaux ont été réalisés sur des corpus bruités[3] qui soulèvent des problèmes similaires à ceux posés par notre corpus. Nous nous intéresserons donc aux méthodes employées pour traiter ces corpus et qui pourraient nous aider dans le traitement de notre corpus.

3.1. PRÉSENTATION DU TAL

Le but du TAL est la « conception de logiciels (programmes) capables de traiter de façon automatique des données linguistiques, exprimées en langues "naturelles" » (Fuchs, 1993, p.7). Les langues dites naturelles font référence aux langues parlées ou écrites par les humains (Bouillon, 1998) et sont opposées aux langages artificiels utilisés dans des domaines comme l'informatique, les mathématiques et la logique. La présentation qui suit se base sur deux livres majeurs, celui de C. Fuchs et A. Lacheret-Dujour, B. Victorri, L. Danlos, et D. Luzzati (1993) et celui de P. Bouillon (1998).

[3] Nous nous appuyons sur l'usage qu'a M. Baranes (2012) de ce terme. Un texte bruité est un texte présentant un grand écart à la norme.

3.1.1. ORIGINES DU TAL

À l'origine, le TAL se limitait au champ de la traduction automatique, initié dans les années 1950, dans un contexte d'après-guerre et de guerre froide. La traduction automatique était alors vue comme un simple exercice de décryptage et devait profiter à l'armée et aux services d'espionnage. Mais très vite, l'illusion tombe et en 1966, le rapport ALPAC affirme la non rentabilité de ces recherches. Les financements sont immédiatement arrêtés et les spécialistes du domaine sont obligés de diversifier leurs recherches. Ce qui conduit, dans les années 1970, à l'émergence ou au développement de nombreux domaines du TAL comme le dialogue Homme-Machine, la synthèse vocale, la correction, etc. À partir des années 1975, c'est l'essor du TAL. Cette discipline prend de plus en plus d'importance, notamment grâce au développement de l'informatique, à l'émergence de nouvelles théories linguistiques et à une masse d'informations de plus en plus croissante. Le TAL va permettre de traiter de plus en plus d'informations sur des corpus de plus en plus grands.

3.1.2. NIVEAUX D'ANALYSE

On distingue traditionnellement deux types de traitement en TAL : l'analyse et la génération. L'analyse permet de passer d'un texte en langue naturelle à une représentation en langage artificiel, utilisable par la machine. En termes d'objectifs, la génération peut être considérée comme l'opération inverse de l'analyse. Elle permet d'obtenir en sortie un texte (oral ou écrit) compréhensible par l'homme à partir d'une représentation formelle. Cette opération peut nécessiter une étape préalable permettant de transformer des données en langage interprétable par l'ordinateur, au besoin. Selon les applications, ces deux types de traitement peuvent être nécessaires ou non. Ainsi, en dialogue Homme-Machine, il est nécessaire d'analyser la question de l'humain pour générer une réponse, tandis qu'en production automatique de bulletins météorologiques, seule la génération est nécessaire. Notre étude a pour but l'élaboration d'un outil d'annotation qui prend en entrée des productions écrites en langage naturel afin de les transformer en données annotées. Notre outil est donc clairement fondé sur des techniques d'analyse automatique.

Chacun de ces traitements traverse différents niveaux linguistiques dans un ordre différent. Les auteurs ne s'accordent pas toujours sur le nombre de ces niveaux – ni même sur leur enchainement -, qui peuvent dépendre du traitement et de la modalité, orale ou écrite, de l'application. Sans entrer dans ce débat, nous pouvons tout de même évoquer les principaux niveaux.

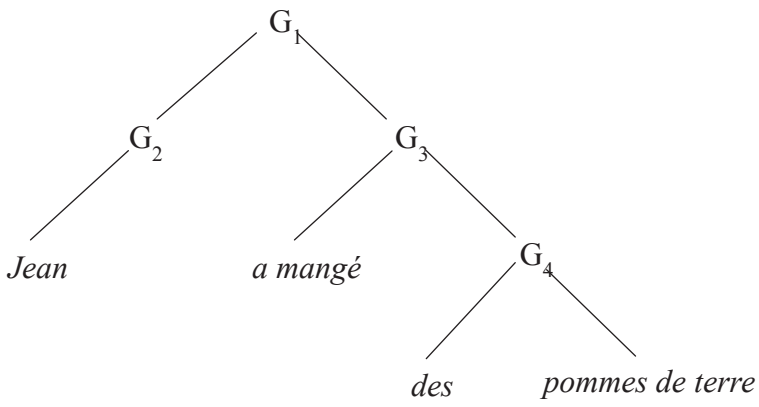
– Le niveau **phonétique et phonologique** permet d'étudier la façon dont les sons se combinent pour former des mots. Ce niveau est particulièrement utilisé en traitement de la parole. En synthèse vocale, il permet de transcrire un texte en phonèmes (unité fonctionnelle de l'oral), c'est l'étape de phonétisation (niveau phonologique). Puis ces phonèmes sont transformés en sons (niveau phonétique). À l'inverse, en reconnaissance vocale, il s'agit de transformer une chaine de parole continue en phonèmes puis en mots.

– Le niveau **prosodique** permet de s'intéresser à la segmentation de l'oral en unités d'expression ou de rythme. Comme précédemment, ce niveau est principalement utilisé en synthèse et en reconnaissance vocales. Il permet, en synthèse, de produire une parole perçue comme plus « naturelle ». Il facilite également la reconnaissance vocale en segmentant la parole en unités relativement larges.

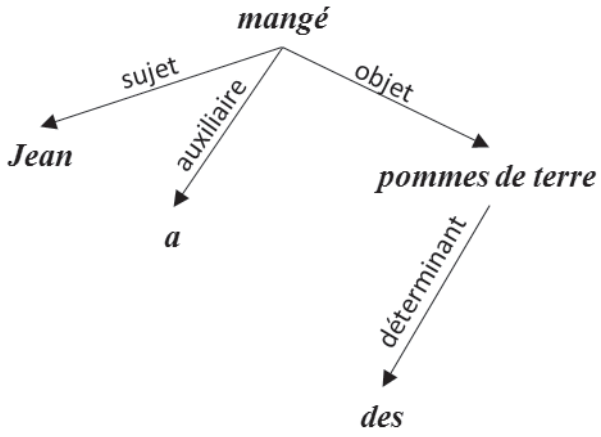
– Le niveau **lexical et morphologique** porte sur la façon dont les morphèmes s'assemblent pour former des mots. Les morphèmes sont des unités de signification qui peuvent correspondre à des affixes, des radicaux ou des mots entiers. Le mot *impensable* se décompose en trois morphèmes : *im-*, *pens*, *-able* (exemple tiré de Bouillon, 1998, p. 13). Ce niveau permet également d'attribuer des informations morpho-syntaxiques aux mots, comme la catégorie syntaxiques, le genre, le nombre, etc. et des attributs sémantique, comme le fait d'être animé ou non.

– Le niveau **syntaxique** permet de s'intéresser à la façon dont les mots se combinent en phrases. Actuellement, on distingue principalement deux approches d'analyse syntaxique. L'analyse syntaxique en constituants considère que les unités se rassemblent en groupes syntaxiques, ou syntagmes, comme les groupes nominaux, les groupes prépositionnels, etc. Il permet ainsi de fournir en sortie une représentation des groupements syntaxiques qui régit la phrase. Pour la phrase *Jean a mangé des pommes de terre* (exemple tiré de Fuchs, 1993, p.106), on peut obtenir les représentations :

$$[[\underline{Jean}_{u1}]_{G2}\ [\underline{a}\ \underline{mangé}_{u2}\ [\underline{des}_{u3}\ \underline{pommes\ de\ terre}_{u4}]_{G4}]_{G3}]_{G1}$$



L'analyse syntaxique en dépendance envisage plutôt la syntaxe comme des relations entre des têtes lexicales et les arguments qu'elles déterminent. Ainsi, une tête nominale peut sélectionner un déterminant comme premier argument et un complément de nom en deuxième argument. Pour l'exemple précédent, on obtient la représentation en graphe de dépendances suivante :



– Le niveau **sémantique** porte sur le sens des énoncés. Il permet d'étudier la façon dont les sens des mots se combinent pour donner un sens global à la phrase. Cette étude se fait généralement par attribution de traits sémantiques aux « mots pleins » (noms, adjectifs, verbes). Analyser une phrase revient alors à analyser la combinaison de ces traits. La représentation des concepts et de leurs associations sémantiques prend généralement la forme de réseaux sémantiques.

– Le niveau **pragmatique** permet d'analyser une phrase par rapport à son contexte d'énonciation. Il met en relation la phrase énoncée, le contexte d'énonciation et les connaissances du monde, afin de rendre la phrase plausible ou de juger de sa plausibilité.

Afin de traiter les erreurs d'orthographe et de segmentation, notre étude sera essentiellement portée sur le plan lexical. Cependant, des analyses sur les niveaux phonologique et syntaxique seront également utiles.

Il est d'usage en TAL de distinguer les applications qui relèvent de l'oral de celles qui relèvent de l'écrit. Toutefois, selon C. Fuchs (1993), l'objectif à terme est de faire se rejoindre ces deux modalités, comme c'est déjà le cas dans certaines applications à l'exemple du dialogue Homme-Machine. De plus, certains outils peuvent s'apparenter à l'écrit tout en empruntant des méthodes à l'oral, et inversement, comme c'est le cas dans notre application. En effet, notre corpus est un corpus entièrement manuscrit, il relève donc de l'écrit. Cependant, comme nous l'avons vu dans le chapitre précédent, l'oral y occupe une place importante, nous ne laisserons donc pas de côté les hypothèses de l'oral.

3.2. SEGMENTATION ET TOKENISATION

Afin de normaliser un corpus en langue française, il est nécessaire de repérer les unités significatives du texte et de les segmenter puis d'en normaliser l'orthographe. Cette segmentation peut se faire à différents niveaux d'analyse. Ainsi, l'on distingue :

- la segmentation morphologique, qui permet de distinguer les morphèmes ;
- la segmentation lexicale, qui permet de repérer les mots, appelés tokens en TAL ;
- la segmentation syntaxique, qui permet de segmenter en syntagmes, en chunks (unité syntaxique basée sur une perception prosodique et regroupée autour d'une tête lexicale, Abney, 1991), en propositions ou encore en phrases.

Dans un but de normalisation, la segmentation porte principalement sur le mot, il s'agit d'une segmentation lexicale, appelée tokenisation. Mais lorsque la tokenisation est troublée par différents facteurs, il peut être nécessaire d'avoir recours à une segmentation en unités plus larges comme la proposition ou la phrase. Ces différentes segmentations sont envisagées de manières très diverses selon les langues et les types de corpus.

3.2.1. TOKENISER EN S'APPUYANT SUR LES FRONTIÈRES DE MOTS

D. D. Palmer (2000) distingue la tokenisation de la segmentation en mots. La tokenisation est le processus de séparation des séquences de caractères en localisant les limites entre les mots. La segmentation en mots vise le même résultat mais sans qu'il y ait de frontière de mots apparente. En français, les frontières de mots sont le plus souvent apparentes et sont marquées par des espaces. Ainsi, la phrase « Le petit chat avance. » (**588**, *Le petit chat avance.*) donnera lieu à la segmentation «<Le><petit><chat><avance>». Cependant, l'espace n'est pas un marqueur entièrement fiable, c'est le cas dans l'exemple *pomme de terre* qui doit être considéré comme un seul token <*pomme de terre*>. L'espace n'a donc pas pour fonction, ici, d'être séparateur de mots. De plus, l'on peut trouver parfois d'autres séparateurs comme dans *c'est* ou encore *Est-il là ?*, où l'apostrophe et le tiret ont remplacé l'espace.

De même, une analyse identique peut être faite au niveau de la phrase et de ses signes de ponctuation, comme le point, le point d'interrogation, le point d'exclamation et les points de suspension, qui ne sont pas exempts de toute ambigüité. En effet, le point est principalement un signe de fin de phrase, mais il peut également être une marque de fin d'abréviation, comme dans « etc. », se placer dans un nombre, ou marquer un changement de ligne lorsqu'il s'agit d'un corpus d'enfants.

Selon C. Fuchs (1993), une solution possible pour tokeniser un texte tout en considérant les ambigüités citées plus haut est d'utiliser deux ressources. D'une part, il est possible d'utiliser des listes fermées d'« exceptions » pour traiter, par exemple, les mots incluant une apostrophe, comme *aujourd'hui*. D'autre part, un système de règles permettrait de désambigüiser le tiret et le point notamment à l'aide du contexte, comme dans l'exemple *Est-il là ?*. Ces règles peuvent s'appuyer sur des schémas syntaxiques, morphologiques ou encore sur des indices comme la casse et la catégorie grammaticale d'un mot pour la segmentation en phrase. Une fois énoncées, ces règles peuvent être transcrites en expressions régulières (Mourad, 2001 ; Mikheev, 2003). D'autres travaux ont contribué à étoffer les règles permettant de désambigüiser les points contenus dans les abréviations (Grefenstette et Tapanainen, 1994) et les noms propres (Mikheev, 2002), mais ces cas ne concernant pas notre corpus, nous ne nous y attarderons pas.

La désambigüisation de ces marqueurs peut également être vue comme un problème de classification (Schmid, 2007). L'on peut alors considérer plusieurs classes, notamment la classe « frontière de mot », la classification permettant de déterminer si le signe ambigu appartient à cette classe ou non. Différents algorithmes ont été testés pour cette méthode, comme les arbres de décision (Riley, 1989), les transducteurs (Silberztein, 1993), les réseaux de neurones (Palmer et Hearst, 1997 ; Mikheev, 2000) et une approche à entropie maximale (Reynar et Ratnaparkhi, 1997). Cependant, cette approche nécessite un corpus d'entrainement annoté à la main. Afin d'éviter ce cout, des méthodes d'apprentissage non supervisées ont également été testées (Grefenstette et Tapanainen, 1994 ; Mikheev, 2002).

Notre corpus étant en français, des analyseurs de ce type peuvent être employés pour une grande partie de nos textes. Toutefois, la maitrise de la segmentation en mots n'est pas toujours acquise aux CP. En effet, notre corpus présente des textes ou des portions de texte mal segmentés voir non segmentés.

Exemples :

Production respectant la segmentation normée : « Le petit chaton marche sur un / marche. Le petit chaton tonbe sus / la marche. Le petit chaton pler / et sa maman chat se reveill. La / maman chat pran le petit chaton / des sa bouche. » (***1345***, *Le petit chaton marche sur une marche. Le petit chaton tombe sur la marche. Le petit chaton pleure et sa maman se réveille. La maman chat prend le petit chaton dans sa bouche.)*.

Production mal segmentée : « . le chasva de se dre la qualine / . é li teno dé sur la tépé e a pe / . re li pelire é la maman / . le ré que péré lé chas. » (**3071**, *Le chat s'en va <incomprehensible>. Et il tombe sur la tête et après il pleure et la maman le récupère le chat.*).

Production sans marque de segmentation : « lechaineMil/ilchataitonbé/lechaserenaine » (**3028**, *Le chat il miaule. Le chat est tombé. Le chat se promène.*).

Il nous faut alors envisager des méthodes ne se basant pas exclusivement sur les marqueurs de frontière de mots.

3.2.2. SEGMENTATION EN L'ABSENCE DE MARQUEURS

La segmentation de textes sans marqueurs de frontière a été envisagée à partir de deux angles d'approche : le point de vue adopté pour les langues ne présentant aucune frontière de mots comme le japonais et le chinois et le point de vue adopté en reconnaissance vocale et pour les transcriptions de parole où aucune frontière de mot n'est audible ou visible.

Schmid (2007) distingue quatre approches principales pour traiter les langues sans frontières de mots. La première de ces approches est la méthode par règles (Ma et Chen, 2003) qui utilise un dictionnaire de mots et un ensemble de règles. À partir du dictionnaire, le texte est divisé en séquences de mots potentielles, la désambigüisation se fait au moyen des règles. Une seconde approche proposée est basée sur les systèmes statistiques qui définissent la tâche de tokenisation comme celle de trouver la séquence de mots la plus probable, notamment au moyen des probabilités conditionnelles.

Ces deux approches font appel à des dictionnaires pour identifier les mots du texte. Ceci présuppose que le texte soit correctement orthographié. Or dans notre corpus, ce n'est souvent pas le cas, prenons l'exemple cité plus haut :

« lechaineMil/ilchataitonbé/lechaserenaine » (**3028**, *Le chat il miaule. Le chat est tombé. Le chat se promène.*) On a ici la séquence « ilchataitonbé » à segmenter en « <il><chat><ai><tonbé> ». Dans ce cas, ces systèmes risquent d'échouer devant « tonbé », forme non existante dans un dictionnaire.

La troisième approche permet de se soustraire à la nécessité d'employer un dictionnaire, il s'agit de l'approche basée sur un étiquetage. Elle permet d'indiquer, à l'aide d'une étiquette, si un caractère donné commence un mot ou non (Fu et Luke, 2003), l'étiquette étant établie à partir de calculs statistiques ou de vecteurs. Contrairement à G. Fu et K.-K. Luke qui n'utilisent que deux étiquettes, N. Xue et L. Shen (2003) utilisent quatre étiquettes différentes, selon que le caractère est en début, milieu ou fin de mot ou qu'il est constitué d'un caractère isolé.

Toutefois, cette approche nécessite que les débuts et fins de mots soient bien orthographiés, ce qui n'est pas non plus toujours le cas dans notre corpus :

« Listoir du peti cha » (**3129**, *L'histoire du petit chat*)

Dans cet exemple, beaucoup de fins de mots sont supprimées, de même que certains débuts de mots, comme la lettre *h* de *histoire*.

Enfin, la dernière approche proposée intègre un étiqueteur morphologique (Jiang, Liu, Chen et Lu, 2004) ou une analyse syntaxique (Wu, 2003). Ces approches ressemblent aux étiqueteurs morphologiques et analyseurs standards, excepté que les débuts et fins de mots n'étant pas fixés, le nombre de possibilités est plus important. La désambigüisation des tokens doit alors être faite en même temps que la désambigüisation de l'analyse.

Comme nous l'avons déjà évoqué, chacune de ces analyses se confronte à un problème de taille : l'éloignement de notre corpus à la norme orthographique. Étant donné le nombre important d'erreurs n'altérant pas la valeur phonique dans notre corpus (cf. 2.4.), un moyen de contourner ce problème serait d'utiliser la forme phonique et non la forme graphique. Le texte se présente alors comme une suite de phonèmes à l'identique de ce que l'on retrouve en reconnaissance vocale, une fois le signal transcrit en phonème.

L'approche principale des systèmes de reconnaissance vocale actuels présente une architecture composée de deux modèles stochastiques : un modèle acoustique et un modèle de langage (Haton, *et al.*, 2006). Les modèles acoustiques sont composés des éléments que l'on souhaite reconnaitre, en l'occurrence les mots. On trouvera donc dans ces modèles un lexique contenant les formes phonétiques des mots. Différentes méthodes peuvent être utilisées pour l'implémenter, notamment le modèle de Markov caché qui permet un apprentissage automatique et le modèle connexionniste qui simule le fonctionnement humain. Ces modèles permettent de reconnaitre différents mots candidats. Les modèles de langage contiennent les connaissances syntaxiques et sémantiques qui permettent de discriminer les mots candidats. Composés à l'origine d'automates à états finis ou de grammaires hors contextes, les méthodes statistiques sont maintenant les plus utilisées, à l'exemple des modèles *n-grammes*.

3.3. NORMALISATION ORTHOGRAPHIQUE

Les différentes méthodes de tokenisation s'appuyant pour la plupart sur des tokens et des connaissances lexicales, il est nécessaire que les tokens étudiés approchent la norme. Or, le corpus que nous nous proposons de traiter en est relativement éloigné. Cette réalité se traduit dans la segmentation en mots mais elle est d'autant plus visible au niveau de la

correction lexicale. Une phase de détection des erreurs et de correction de celles-ci en vue de normaliser les unités lexicales sera donc nécessaire. Dans cette optique, seule la correction lexicale nous a paru intéressante dans un premier temps. Pour un retour sur le développement des correcteurs grammaticaux et de leur fonctionnement, il est possible de se référer à Vienney et Bioud (2004).

3.3.1. CLASSIFICATIONS DES ERREURS

De nombreuses typologies d'erreurs lexicales ont été avancées au cours des travaux de détection et de correction, adaptées au besoin de chaque système. Les premiers systèmes distinguaient quatre types d'erreurs (Levenshtein, 1966) : l'addition, le déplacement, l'omission et la substitution. Puis, ces modèles ont évolué et se sont étoffés.

Certains systèmes actuels différencient les erreurs de compétence (Véronis, 1988), qui correspondent à une mauvaise connaissance de la langue, des erreurs de performance, dues à l'emploi de la langue en contexte et principalement l'utilisation de l'outil informatique (substitution par une lettre proche au clavier, par exemple). De manière quasi-similaire, le système VORTEX (Pérennou, 1990) différencie erreurs linguistiques et erreurs typographiques. Les erreurs linguistiques sont entendues dans un sens large et englobent toutes les erreurs intervenant dans la phase de conceptualisation du texte. Il peut s'agir d'erreurs orthographiques, stylistiques, etc. Les erreurs typographiques correspondent à la phase de transcription matérielle du texte mental. Elles incluent les erreurs de segmentation et de disposition du texte, ou encore les erreurs d'insertion, d'effacement ou de substitution des lettres.

Cependant, les erreurs typographiques évoquées dans ces classifications correspondent bien souvent à des erreurs dues à l'utilisation du clavier. Or, notre corpus est un corpus manuscrit, cette classification n'est donc pas pertinente. De plus, la majorité des erreurs décrites dans cette catégorie sont observées dans notre corpus mais correspondent, pour des scripteurs débutants, à des erreurs de compétence.

Dans le système DECOR, V. L. Strube De Lima propose une classification en trois catégories :

– Les erreurs typographiques, catégorie qui regroupe également les erreurs orthographiques.

– Les erreurs phonétiques basées sur les correspondances entre la prononciation et l'orthographe. Appelées également erreurs phonogrammiques (Catach, 1984) ou

phonographiques (Ghneim, 1997), elles correspondent aux erreurs pour lesquelles les formes orales d'une forme erronée et de sa forme normée sont identiques ou proches, par exemple **estauma* / *estomac*. Comme nous l'a montré notre première observation, ces erreurs sont nombreuses dans notre corpus.

– Les erreurs de génération qui englobent les phénomènes dus à une méconnaissance de certains éléments grammaticaux. Par exemple : **chevals* / *chevaux.*

Certains systèmes se basent également sur la typologie proposée par N. Catach, D. Duprez et M. Legris (1980), distinguant erreurs extragraphiques, dont les erreurs phonétiques (phonie altérée) et erreurs graphiques (phonie non altérée). Cette classification fera l'objet d'un exposé plus détaillé ultérieurement (cf. chapitre 5).

Les différentes classifications présentées ici pourront nous inspirer mais ne peuvent en l'état être adoptées pour notre corpus. Celui-ci présente, en effet, des caractéristiques majeures différentes des corpus pour lesquels elles ont été établies, entre autres son caractère manuscrit.

3.3.2. MÉTHODES DE CORRECTION

La plupart des systèmes de correction ont été élaborés pour des scripteurs experts écrivant dans leur langue maternelle et ne produisant qu'accidentellement des erreurs d'orthographe (Heift et Schulze, 2007). Dans ces écrits, les erreurs y sont non systématiques et peu fréquentes. Ces correcteurs ont donc été conçus pour des productions en langue native très proche de la norme orthographique.

3.3.2.1. APPROCHES PAR LEXIQUES

Les premiers correcteurs orthographiques reposaient sur des analyses hors-contextes où le mot était considéré indépendamment de ses voisins (Baranes et Sagot, 2014). Ces correcteurs se basaient essentiellement sur des systèmes de règles typographiques et sur un calcul des distances d'édition proposées par V. Levenshtein (Damerau, 1964 ; Kernighan, Church et Gale, 1990). Ce calcul prend en compte quatre types d'erreur : l'insertion, la suppression, la substitution et le déplacement. Cependant, cette approche est relativement peu efficace pour les mots comportant plus d'une erreur ou pour les mots pour lesquelles plusieurs formes normées peuvent être proposées (Baranes, 2012).

Correction par clé de similarité

Afin de pouvoir traiter des mots comportant un plus grand nombre d'erreurs, des méthodes à base de comparaison de clé sont apparues. Les clés sont des chaines de caractères calculées à partir des lettres des mots. Après calcul de la clé de la forme erronée, elle est comparée à un lexique de clés, ce qui permet d'obtenir les graphies les plus proches. Ce type de traitement permet principalement de traiter les doublements ou les suppressions de lettres et les erreurs d'accentuation.

Pollock et Zamora (1984) proposent ainsi une comparaison par clé squelette. Cette clé est calculée en conservant la première lettre du mot puis en concaténant les consonnes dans l'ordre d'apparition et en supprimant les doublons, de même pour les voyelles, les accents étant supprimés au préalable. Pour les mots *préférer* et **prèferrer* on obtient ainsi la clé *pfre*.

En se basant sur le constat que le français est une langue où les anagrammes sont rares, Debili et son équipe (1986) ont proposé une clé appelée alphagramme calculée en concaténant les lettres du mot dans l'ordre alphabétique. *Préférer* et **prèferrer* ont pour clé respectivement *eééfprrr* et *eeèfprrrr*.

Enfin, M. Ndiaye et A. Vandeventer Faltin (2004) proposent un alphacode proche des clés squelettes qui se calcule en concaténant les consonnes ordonnées par ordre alphabétique aux voyelles ordonnées de même. Les accents et les doublons étant omis. *Préférer* et **prèferrer* ont pour clé *fpre*.

Toutefois, pour certaines erreurs, comme les erreurs phonogrammiques et phonétiques, les graphies entre forme erronée et forme corrigée peuvent être assez éloignées. Pour traiter ces erreurs, il est possible d'envisager des clés phonétiques.

Correction par clé phonétique

Parmi les systèmes traitant des erreurs phonétiques, deux types peuvent être distingués : les systèmes qui comparent les représentations phonétiques de la forme erronée et des formes d'un lexique (Laporte et Silberztein, 1989 ; Williams, 1991) et les systèmes qui comparent à un lexique des réécritures graphiques de la forme erronée à partir de correspondances graphèmes-phonèmes (Véronis, 1987 ; Pécatte, 1992 ; Belrhali, 1995).

Les premiers de ces systèmes utilisent des modules de phonétisation qui permettent de transcrire un mot ou un texte en une représentation phonétique. Principalement développés pour des systèmes de synthèse vocale (Divay et Guyomard, 1977 ; Léty, 1980 ; Aubergé,

1985), ils peuvent avoir de nombreuses autres applications comme la correction automatique (Lahens, 1987), la reconnaissance de la parole (Lacheret-Dujour, 1990), ou encore la description linguistique (Catach, 1984). On distingue deux types de systèmes de phonétisation (Ghneim, 1997) :

– Les systèmes à connaissances explicites qui impliquent une description des connaissances linguistiques. À l'origine, ces systèmes étaient constitués de règles de correspondances phonie-graphie. Les exceptions étant prises en compte soit par des listes d'exceptions, soit par des règles spécifiques. Puis, des lexiques ont été ajoutés, cohabitant généralement avec les règles dans les systèmes. Les lexiques peuvent contenir des morphes, des mots et leurs formes fléchies ou dérivées ou encore l'ensemble des formes de la langue. À chaque unité, la forme phonologique est associée. Néanmoins, comme le relèvent N. Torzec, T. Moudenc et F. Emerard (2001) une telle approche atteint rapidement ses limites pour des textes mal formés ou comportant un nombre de mots inconnus trop important. Or, c'est le cas de notre corpus, nous privilégierons donc une approche par règles. Habituellement, ces règles sont organisées en liste par ordre de priorité, des règles les plus spécifiques aux plus générales. Dans un but d'optimisation en limitant les tests inutiles, M. Morel et A. Lacheret-Dujour (2001) ont proposé une organisation des règles sous forme d'arbres.

– Les systèmes à connaissances implicites reposent sur des apprentissages automatiques utilisant des techniques telles que l'apprentissage symbolique (Dietterich, Hild et Bakiri, 1995), les réseaux de neurones (Sejnowski et Rosenberg, 1987) et les modèles Markoviens (Parfitt et Sharman, 1991).

Les seconds systèmes utilisent des tables de correspondances graphèmes-phonèmes. La réécriture des chaines erronées peut se faire soit par des méthodes stochastiques, comme le système VORTEX qui combine règles de correspondance et probabilité, soit par des méthodes déterministes (Véronis, 1987, cité dans Ghneim, 1997), pour lesquelles toutes les possibilités sont développées. Puis, une comparaison à un lexique accompagnée d'un calcul de similarité permet de sélectionner la forme normée.

Correction des erreurs de génération

Enfin, certaines erreurs ne relèvent pas de la phonétique mais de la morphologie, ce sont les erreurs de génération. Pour corriger ce type d'erreurs, le principe adopté par Cohard (1988, cité dans Menézo, 1999) consiste à extraire d'une part le radical, d'autre part les informations contenues dans la désinence erronée, comme le genre et le nombre pour les adjectifs, le temps et la personne pour les verbes. Il s'agit ensuite de fléchir la racine extraite

selon ses informations pour obtenir la forme normée. Par exemple, prenons la forme **chevals*, de laquelle on extrait la racine *cheval* et l'information « pluriel », par la présence du *s*. On cherche alors le pluriel de la racine *cheval*, c'est-à-dire *chevaux*.

Ces modèles reposent sur des lexiques de formes « correctes », que ce soit une forme graphique ou phonique. Seules les formes non contenues dans ces lexiques sont jugées erronées (Jacquet-Pfau, 2001). L'utilisation de telles ressources a alors fait émerger la nécessité de faire une différence entre « non-mot », c'est-à-dire une forme non contenue dans le lexique, et « mot faux », c'est-à-dire une forme erronée du mot qui correspond à un autre mot du lexique (Whitelaw *et al*., 2009). Un « non-mot » est facilement détectable par un lexique mais un « mot faux » l'est beaucoup moins. C'est pourquoi des modèles prenant en compte certaines données contextuelles ont été élaborés.

3.3.2.2. APPROCHES CONTEXTUELLES

La plupart de ces approches se basent sur des modèles de langage *n-grammes* où l'unité est le token (Brill et Moore, 2000 ; Carlson et Fette, 2007 ; Park et Levy, 2011). Cependant, lorsque les contextes gauches et droits du mot erroné sont eux-mêmes erronés, comme c'est souvent le cas dans notre corpus, cette solution risque d'échouer (Baranes, 2012). Une autre solution a alors été proposée où l'unité du n-gramme n'est plus le token graphique mais sa forme phonétique (Toutanova et Moore, 2002). Une autre solution proposée est de considérer ces deux unités, le n-gramme et la forme phonique, dans un même temps (Boyd, 2009).

Certaines approches se sont également appuyées sur le contexte proche à partir de systèmes de règles (Mangu et Brill, 1997). D'autres encore s'appuient sur des analyses distributionnelles des contextes de chaque candidat (Suignard et Kerroua, 2013). Ainsi, le mot erroné sera remplacé par le candidat partageant un contexte similaire.

La majorité de ces approches, cependant, sous-tendent que le mot erroné ne contient qu'une ou deux lettres modifiées au plus, ou bien que le contexte de ce mot n'est pas lui-même erroné, tout au moins dans sa forme phonétique. Ces deux hypothèses s'avèrent souvent vraies pour des scripteurs experts mais plus souvent fausses lorsqu'il s'agit de corpus bruités de scripteurs en cours d'apprentissage.

Les méthodes proposées pour la correction automatique standard n'étant pas adaptées à notre corpus, nous pouvons nous intéresser à la manière dont la correction est envisagée dans le domaine de l'apprentissage des langues où les scripteurs, comme les scripteurs de

notre corpus, ne sont pas experts de la langue et où les productions sont donc plus éloignées de la norme.

3.3.3. ADAPTATION AUX CORPUS PEU NORMÉS

3.3.3.1. CORRECTION D'ERREURS POUR L'APPRENTISSAGE D'UNE LANGUE SECONDE

Dans le domaine de l'apprentissage des langues assisté par ordinateur (désormais ALAO), des travaux ont émergé afin de coupler TAL et ALAO dans le but d'améliorer ces systèmes. Cependant, utiliser des techniques informatisées dans un contexte d'apprentissage nécessite que ces techniques soient fiables et robustes. Les résultats obtenus en correction grâce au TAL ne sont donc pas encore suffisants pour utiliser ces techniques pour toutes les applications (Kraif *et al.*, 2004). Néanmoins, utilisées à bon escient, ces dernières peuvent contribuer à améliorer les systèmes ALAO. De ce fait, les méthodes automatiques de correction orthographique sont souvent utilisées pour traiter des réponses courtes ou semi-fermées (Kraif, 2005). En effet, de telles réponses permettent de s'appuyer sur le contexte et sur la réponse attendue pour l'analyse des erreurs. La comparaison entre réponse attendue et réponse obtenue peut se faire avec une utilisation de l'outil TAL en quantité variable. Ainsi, les traitements peuvent s'apparenter à une légère amélioration du pattern-matching acceptant une différence de lettre au plus, à l'exemple du système CAMILLE (Chanier, 1996). Mais ils peuvent également inclure différents niveaux d'analyse (Kraif, 2005), comme la normalisation graphique et les différences orthographique, morphosyntaxique et lexicosémantique, qui à l'aide d'heuristiques de différence permettent de détecter les erreurs. D'autres recherches incluent également un niveau sémantique permettant de comparer le sens des deux réponses et ainsi d'augmenter le nombre de réponses justes acceptées (Antoniadis *et al.*, 2005).

Toutefois, même si notre corpus présente un contexte restreint, la combinatoire des réponses est beaucoup trop vaste pour pouvoir y appliquer de telles méthodes. Pour l'heure, la plupart des systèmes traitant des questions ouvertes en ALAO emploient des correcteurs orthographiques classiques. Cependant, les correcteurs classiques obtenant des taux d'erreurs non corrigées trop importants, des modules morphologiques et syntaxiques vont leur être ajoutés, à l'exemple de la plateforme Freetext (Granger, Vandeventer et Hamel, 1998). Cette plateforme utilise un correcteur orthographique commercial couplé d'une détection des erreurs syntaxiques par relâchement des contraintes et par identification des homophones.

Notons enfin que certains systèmes utilisent des listes d'erreurs non natives (Mitton 1996) et des systèmes de règles applicables élaborées à la main (Chanier, 1992).

De cet exposé, on peut conclure que peu de travaux ont été réalisés dans ce domaine et que peu de méthodes adaptées à la spécificité de notre corpus ont été proposées. Il nous a alors paru intéressant de nous pencher sur les méthodes utilisées pour d'autres corpus bruités comme les SMS.

3.3.3.2. NORMALISATION DE SMS

Le « Short Message Service » (SMS) est un moyen de communication qui s'écarte souvent des conventions orthographiques (Beaufort *et al.*, 2010). Cet écart à la norme nécessite une phase de normalisation avant tout autre traitement linguistique. Pour l'heure, trois métaphores principales ont été avancées pour réaliser cette normalisation.

Métaphore de la « correction orthographique »

En premier lieu, dans la métaphore de la correction automatique, le SMS est vu comme un texte bruité où le token inconnu de la norme est vu comme une version non conforme de la forme correcte du mot (Kobus, Yvon et Damnati, 2008). En considérant l'hypothèse que la plupart des mots sont correctement orthographiés, l'étape de normalisation ne s'intéresse qu'aux mots absents du vocabulaire et correspond alors à une correction mot-à-mot des tokens hors vocabulaire.

Pour ce faire, E. Guimier de Neef et S. Fessard (2007) proposent un modèle, utilisé dans leur système TiLT, basé sur des lexiques d'abréviations qui permet de systématiser certaines corrections comme **tjs / toujours* (Baranes, 2012). Mais l'approche par correction automatique utilise principalement des méthodes statistiques qui calculent la probabilité de la forme normée au vu de la forme bruitée. Cette probabilité est calculée à partir d'un dictionnaire et de la combinaison de la probabilité de la forme normée f_n et de la probabilité de la forme bruitée f_b sachant la forme normée. La forme normée sélectionnée est la forme qui a la plus grande probabilité conditionnelle (Choudhury, *et al.*, 2007 ; Cook et Stevenson, 2009). Ainsi :

$f_{n\,max}$= arg max $P(f_n|f_b)$

$$= \arg\max \frac{P(fb|fn)P(fn)}{P(fb)}$$

M. Choudhury *et al.* (2007) utilisent ce modèle combiné à un modèle en chaines de Markov basées sur des chaines de caractères pour prendre en compte à la fois la forme

graphique et la forme phonémique du mot. Cette approche permet différentes formes bruitées pour une même forme normée, elle permet également de traiter plus d'une erreur par forme.

Métaphore de la « traduction »

Une deuxième approche est la métaphore orientée traduction (Aw, Zhang, Xiao et Su, 2006) qui voit la normalisation comme une traduction d'une langue source vers une langue cible, la langue source étant le SMS et la langue cible l'écrit normé (Beaufort *et al*., 2010). Cette approche repose sur des modèles entrainés par alignement de textes bruités et de leurs contreparties normalisées.

Cette métaphore permet de gérer les forts écarts à la norme. De plus, elle permet une analyse en groupes de tokens plutôt qu'en tokens (Aw *et al*., 2006). Cependant, cette approche est considérée comme dépassant largement les besoins en normalisation de SMS (Choudhury *et al*., 2007). En effet, la distance entre le SMS et la forme normée est bien moindre qu'en langue étrangère. Cette méthode permet également de gérer des correspondances multiples entre langue source et langue cible, ce qui est bien souvent inutile dans ce contexte où la relation est souvent unique. De plus, cette approche repose sur l'automaticité de certaines correspondances entre forme normée et forme SMS. En effet, les premières sont souvent transcrites de quelques façons différentes seulement. Ainsi, *toujours* est souvent transcrit **tjs* ou **tjrs*. Une telle approche traite donc les formes SMS comme une sorte de norme, même si celle-ci diffère de la norme standard et ne permet donc pas de traiter toute la créativité de la langue utilisée dans les SMS (Kobus *et al.*, 2008). Or notre corpus ne comprend que peu de redondances de correspondances, chaque enfant approchant l'orthographe de manière différente. Une telle métaphore semble donc difficilement adaptable à notre corpus.

Métaphore de la « reconnaissance vocale »

Enfin, la normalisation de SMS a aussi été envisagée par le biais de la reconnaissance vocale (Kobus *et al*., 2008). Cette approche se base sur l'hypothèse que les formes trouvées dans les SMS sont parfois plus proches de leur réalité phonémique que de leur forme graphique. Dans ce modèle, le SMS est vu comme une transcription possible d'une forme phonique. La première étape va donc être de retrouver la suite phonique encodée. Les méthodes utilisées en reconnaissance vocale peuvent alors être appliquées. Les phones donnés à l'étape précédente sont découpés en segments de mots puis un modèle de langage est utilisé pour sélectionner la séquence la plus probable (Beaufort *et al*., 2010).

Cette approche a pour avantage de permettre de gérer de nombreuses transcriptions d'un même mot et de ne pas être influencée par des frontières de mots bruitées. Néanmoins, elle ne permet pas de tenir compte des graphèmes initiaux. De plus l'application de cette approche nécessite que tous les phonèmes soient encodés, or de nombreux corpus bruités, et notamment le nôtre, présentent des élisions, comme l'absence de voyelles dans certains mots. Par la suite, des approches combinant plusieurs de ces méthodes ont été élaborées. C'est le cas notamment de l'approche de R. Beaufort *et al*. (2010) qui mêle traduction et correction.

Le but de ce mémoire est de poser les bases de l'annotation d'erreurs d'un corpus bruité, à savoir des productions écrites d'élèves de CP, à partir de méthodes du traitement automatique des langues. Cependant, peu de méthodes parcourues dans cet état de l'art ne peuvent être appliquées en l'état à notre corpus, il va donc nous falloir élaborer des méthodes spécifiques répondant aux exigences de notre corpus.

CHAPITRE 4 - L'ORTHOGRAPHE DU FRANÇAIS

Afin de mener à bien la description de l'orthographe du français, il nous faut au préalable différencier l'écriture de l'orthographe. L'écriture correspond aux ressources graphiques mises à disposition des scripteurs d'une langue pour représenter les faits linguistiques afin de transmettre un message dans le temps et dans l'espace. L'orthographe est la norme graphique et sociale établie en sélectionnant et fixant une représentation de la langue parmi ces ressources. Selon les termes de J.-P. Jaffré (Fayol et Jaffré, 2008, p. 28), « l'écriture propose, l'orthographe dispose ».

4.1. L'HISTOIRE DU SYSTÈME D'ÉCRITURE FRANÇAIS

Les premières écritures apparues en Mésopotamie au IV^{e} millénaire av. J.-C. n'ont pas pour but de transcrire phonétiquement la langue et notent majoritairement des noms. Ces types d'écritures sont qualifiés d'écritures idéographiques ou logographiques (Fayol et Jaffré, 2008). Au fur et à mesure de leur évolution, ces systèmes vont se doter de procédés phonographiques plus économiques. Ces procédés sont basés sur la relation entre oral et écrit et vont se constituer selon deux types d'unités possibles : la syllabe ou le phonème. Cependant, afin que le principe phonographique soit économique, il est nécessaire que le nombre d'unités ne soit pas trop élevé. Dans de nombreuses langues, l'inventaire des phonèmes s'élève à quelques dizaines d'unités tandis que celui des syllabes s'élève à quelques milliers. Seules les langues dont le nombre de syllabes se compte en centaines ont pu adopter un principe syllabique. C'est pourquoi, la grande majorité des langues du monde, à l'instar du français, ont opté pour un principe alphabétique.

En tant que système d'écriture alphabétique, l'écriture française repose sur des procédés phonographiques. Fayol et Jaffré définissent la phonographie comme suit : « ensemble des procédés qui, dans une écriture – ou dans une orthographe-, permettent d'établir des correspondances entre des unités graphiques et les unités phoniques que sont les phonèmes ou les syllabes » (2008, p. 232). Mais tout n'est pas si simple, et pour cause, l'orthographe du français est parfois considérée comme l'une des plus difficiles au monde.

En effet, si 96% des correspondances graphèmes-phonèmes sont régulières, seulement 71% des correspondances phonèmes-graphèmes le sont (Fayol et Jaffré, 2008). Cela revient à dire que pour une graphie inconnue donnée, il est possible de prédire à 96% sa forme oralisée, tandis que pour une forme sonore donnée il n'est possible de prédire sa graphie qu'à 71%. Ce

constat s'explique par le fait que pour une unité graphique donnée, il peut exister plusieurs correspondants phoniques, mais surtout que pour un phonème donné, il existe souvent de nombreux correspondants graphiques. Les auteurs appellent cela la « polyvalence phonographique » (Fayol et Jaffré, 2008, p. 89). Chaque langue présente un degré de transparence différent, certaines études proposent une typologie des langues à orthographes latines (tableau 9).

Degrés	Orthographes
1	anglais
2	**français**, danois
3	allemand, suédois, norvégien, islandais, grec
4	espagnol, hongrois
5	finnois

Tableau 9. Classement des orthographes selon leur transparence (Fayol et Jaffré, 2008, p. 89)

Une orthographe est dite transparente lorsque les correspondances phonographiques qui la composent sont régulières, c'est-à-dire qu'il y a bijection entre phonèmes et graphèmes : un unique graphème transcrit un unique phonème et inversement. Le degré 5 étant le degré de transparence le plus élevé, le français est jugé comme une langue peu transparente. Les correspondances phonographiques du français sont donc irrégulières, un phonème n'est pas transcrit par une seule graphie et une même graphie peut transcrire plusieurs phonèmes. On peut dire que le français est une langue à forte complexité phonographique.

Plusieurs raisons peuvent expliquer cette complexité. On peut notamment évoquer des raisons historiques. En effet, pour écrire le français, c'est l'alphabet latin qui a été choisi. Cependant le latin est, par certains égards, moins riche phonologiquement que le français. Certains signes hérités du latin ont donc été utilisés pour transcrire en français deux phonèmes différents. Certaines stratégies orthographiques ont alors été développées pour distinguer les différents usages d'un même signe. Pour exemple, le signe *u* permettait à la fois de désigner le phonème /v/ et le phonème /y/. Sur certains mots un *h* est apparu pour marquer l'usage du *u* en tant que correspondant du /y/, comme dans *huitre*, *huit*, etc. (Catach, 1978). Un tel procédé permet de distinguer des mots auparavant homographes, comme *huitre* et *vitre* ou *huit* et *vit*.

L'évolution plus rapide de l'oral par rapport à l'écrit constitue un autre facteur de complexité. En effet, si notre orthographe contient beaucoup d'oppositions *o* / *au*, par exemple, c'est parce qu'il y a eu une différence de prononciation correspondant à ces deux graphies (Fayol et Jaffré, 2008). Désormais, la prononciation a changé mais l'orthographe n'a pas accompagné ce changement.

Plus encore, cette polyvalence orthographique s'explique par la tendance qu'a l'écrit à différencier les homophones. Dans le cadre du français, cette tendance s'est accrue entre le XII^{e} s. et le XIV^{e} s. avec le développement de la lecture silencieuse. Pour le lecteur, il s'agissait alors de diminuer le besoin d'indices oraux, ce qui a eu pour effet d'augmenter les indices visuels afin de faciliter la lecture qui n'est plus un décodage sonore mais visuel (Fayol et Jaffré, 2008). L'orthographe acquiert ainsi un caractère sémiographique, tout en conservant sa base phonographique.

4.2. DÉCRIRE L'ORTHOGRAPHE

4.2.1. LE GRAPHÈME

Afin d'étudier le système orthographique et son fonctionnement, il nous faut adopter une unité d'étude, le graphème. Si l'utilisation de cette unité fait relativement consensus dans la communauté scientifique, ce n'est pas le cas de sa définition. Dès 1979, N. Catach synthétise les problèmes et les points de débats entre les auteurs. Plus tard, J.-C. Pellat (1988) propose une revue des évolutions de la notion de graphème et distingue quatre ensembles de définition pour cette seule notion.

Le premier ensemble de définitions envisage le graphème comme l'unité minimale du code écrit (Stetson, 1937 ; Haas, 1976, cités dans Pellat, 1988). Le graphème est alors souvent confondu avec la lettre, à l'image de C. Blanche-Benveniste et A. Chervel pour qui « la distinction entre le graphème et la lettre se fonde sur des considérations identiques à celles qu'utilisent les phonologues pour opposer le phonème au son » (1969, p.119). Le graphème est donc une unité abstraite tandis que la lettre en est la réalisation en contexte. Dans cette définition, le graphème est défini par des caractéristiques purement graphiques, ce qui n'empêche pas les auteurs d'en faire une analyse basée sur une dimension phonique (cf. 4.2.2.).

Un deuxième ensemble de définitions, continuant de considérer le graphème comme l'unité minimale de l'écrit, lui confère un caractère indépendant de l'oral. Ces définitions lui attribuent alors une dimension sémiographique. L. Hjelmslev (1957, cité dans Pellat, 1988) distingue ainsi les plérématèmes ou plérèmes, graphèmes correspondant à des morphèmes (*t* dans *cha<u>t</u>*), des cénèmatèmes ou cénèmes, graphèmes vides de sens et transcrivant des phonèmes (*a* dans *ch<u>a</u>t*). Il est toutefois regrettable que peu de définitions décrivent en substance le graphème et la plupart des auteurs le définissent par analogie au phonème (Uldall, 1944 ; Pulgram, 1951 ; Gleason, 1961).

Un troisième ensemble de définitions considère, à l'opposé, le graphème comme le quasi pendant écrit du phonème (Imbs, 1971 ; Horejsi, 1972 ; Gak, 1976). Il ne peut donc se définir que par référence à celui-ci et peut comporter une ou plusieurs lettres. R. Jakobson (1976, p.77) donne la définition suivante du graphème : « La raison d'être du graphème bêta consiste à désigner le phonème /b/, et tout autre graphème remplit une tâche similaire. L'image graphique fonctionne comme signifiant et le phonème comme son signifié ».

Enfin, un dernier ensemble de définitions repose sur la synthèse de ces définitions, ce qui fait du graphème une unité polyvalente et fonctionnelle qui n'a plus pour seule valeur la valeur phonique. On retrouvera ainsi la définition donnée par M. Fayol et J.-P. Jaffré (2008, p. 230) : « Le graphème est la plus petite unité fonctionnelle de l'écriture. Il peut correspondre à des phonèmes ou à des syllabes. Ainsi le graphème *in* correspond au phonème /ɛ̃/. Le graphème peut également jouer un rôle plus spécifique, notamment grammatical, comme avec le *s* du pluriel français ».

Afin de clarifier la distinction entre graphème et lettre, les auteurs ajoutent également « Les lettres, comme les accents, servent à construire les graphèmes mais ne se confondent pas avec eux ».

N. Catach (1979) reprend les notions de cénèmes et plérèmes et attribue une double nature aux graphèmes. En effet, ils peuvent être à la fois cénémiques, c'est-à-dire qu'ils renvoient à un correspondant phonique, mais ils peuvent également être plérémiques, lorsqu'ils renvoient à un signifié de manière directe, c'est-à-dire lorsqu'ils portent un sens indépendamment de l'oral. L'auteur (1980, p. 16) donnera ainsi la définition suivante du graphème : « la plus petite unité distinctive et/ou significative de la chaine écrite, composée d'une lettre, d'un groupe de lettres (digramme, trigramme), d'une lettre accentuée ou pourvue d'un signe auxiliaire, ayant une référence phonique et/ou sémique dans la chaine parlée ».

Toutes ces définitions peuvent être résumées dans le tableau suivant (tableau 10) :

Définition	Nature	Critère	Principaux auteurs
Unité minimale du code écrit	Lettre	graphique	Stetson (1937), Blanche-Benveniste et Chervel (1969), Haas (1976)
Unité significative de l'écrit	Lettre ou suite de lettres	sémiographique	Uldall (1944), Pulgram (1951), Hjemslev (1957), Gleason (1961)
Correspondant graphique du phonème	Lettre ou suite de lettres	phonographique	Imbs (1971), Horejsi (1972), Gak (1976), J Jakobson (1976)
Unité distinctive et significative de l'écrit	Lettre ou suite de lettres	sémiographique et phonographique	Fayol et Jaffré (2008), Catach (1979)

Tableau 10. Synthèse des différences entre les définitions du graphème

C'est principalement sur le dernier ensemble de définitions, qui nous semble prendre en compte tous les aspects de cette notion, que reposera notre travail. Même si, comme nous le verrons par la suite, il nous faudra le reformuler en termes plus fonctionnels.

N. Catach (1995) établit également quatre critères permettant d'identifier une lettre ou une suite de lettres comme étant un graphème, indépendamment de sa fonction.

– Le degré de **fréquence** mesuré en lexique et dans les textes. Ce critère permet d'exclure les séquences atypiques et peu fréquentes que l'on peut rencontrer dans des mots d'origines étrangères par exemple, comme la lettre *ñ* de *cañon*, non considérée comme graphème du français.

– Le degré de **cohésion** et de **stabilité**. Ainsi, la graphie *eau* est considérée comme un graphème car il est stable. En revanche, la graphie permettant de transcrire le son /ɑ̃/ à partir des lettres *e*, *a* et *n* n'est pas considérée comme telle car non stable. En effet, elle peut s'écrire tantôt *ean* (Jean), tantôt *aen* (Caen).

– Le degré de **signifiance** ou de pertinence phonologique. L'auteur donne pour exemple les lettres doubles pour les racines grecques comme *th* et *rh* qui sont phonologiquement stables mais qui ont peu de signifiance puisqu'elles n'ont pas pour fonction de discriminer deux mots : *rhume* ne s'oppose pas à **rume*. Ces suites de lettres ne sont donc pas des graphèmes.

– Le degré de **rentabilité** linguistique ou de créativité linguistique. Une unité est dite rentable si elle peut être dérivée ou fléchie de manière sérielle et créative pour former de

nouveaux mots. À l'exemple du graphème *–er* qui alterne avec *–ère* dans *boulanger* / *boulangère*, mais également dans *boucher* / *bouchère*, *léger* / *légère*, etc.

4.2.2. L'ORTHOGRAPHE : ENTRE SÉMIOGRAPHIE ET PHONOGRAPHIE

La définition du graphème donnée par N. Catach, tout comme celle proposée par M. Fayol et J.-P. Jaffré, attribue deux fonctions au graphème et à l'orthographe de manière générale. L'orthographe du français, issue d'une écriture alphabétique, contient une grande part de phonographie qui permet de transcrire les sons de la langue. Mais, c'est également une orthographe que l'on dit irrégulière et peu transparente, en raison des principes sémiographiques qui la régissent et dont le principal objectif est de présenter les unités linguistiques de la manière la moins ambigüe possible (Fayol et Jaffré, 2008).

J.-P. Jaffré (Fayol et Jaffré, 2008) distingue la sémiographie mineure de la sémiographie majeure qu'il place sur un continuum (Figure 3). La sémiographie est qualifiée de « mineure » lorsqu'elle est entièrement portée par les signes qui sont en lien avec l'oral, c'est-à-dire les phonogrammes, tandis qu'elle est dite « majeure » lorsqu'elle est davantage portée par des signes sans lien avec la phonologie.

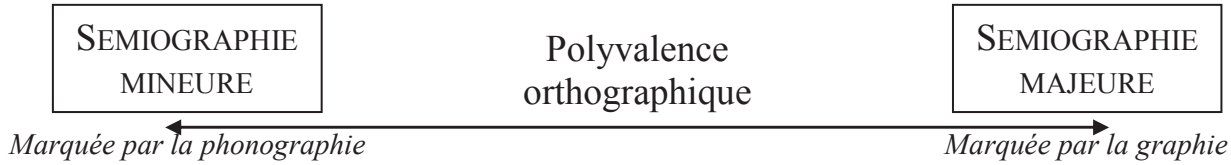


Figure 3. Continuum sémiographique

À titre d'exemple, prenons la forme *tomba*. Dans ce cas, l'écrit n'apporte aucune information supplémentaire par rapport à l'oral, toutes les informations morphologiques sont portées par la phonographie. Les langues dont la sémiographie est majoritairement portée par l'oral sont dites à « potentiel sémiographique élevé » et leur sémiographie est qualifiée de « mineure ».

À l'inverse, en vieillissant, une langue a tendance à produire de plus en plus d'homophonie. Des procédés graphiques émergent alors pour désambigüiser une partie de l'écrit. Ces procédés peuvent correspondre, par exemple, à l'ajout de lettres muettes, comme *s* et *t* qui permettent de distinguer les formes *fais* et *fait*. Les orthographes présentant un grand nombre de ces phénomènes sont qualifiées d'orthographe à sémiographie « majeure ».

En outre, J.-P. Jaffré (Fayol et Jaffré, 2008) place la polyvalence orthographique à mi-chemin entre ces deux procédés. En effet, dans le choix des correspondances phonographiques, la polyvalence orthographique peut porter des indices sémiographiques. Considérons les formes *faim* et *fin*, les graphèmes *aim* et *in*, en tant que phonogrammes, sont impliqués dans la phonologie, mais c'est leurs différences graphiques qui permettent de les distinguer. La sémiographie est donc portée à la fois par la phonographie et par la graphie.

Afin de comprendre à quelles réalités orthographiques renvoient ces différentes fonctions, nous pouvons nous pencher sur les modèles proposés par C. Blanche-Benveniste et A. Chervel (1969) et N. Catach (1978). Bien que ce ne soient pas les seules descriptions réalisées, ces modèles sont encore très présents, notamment dans les recherches en didactique.

4.2.2.1. LE FONCTIONNEMENT DE L'ORTHOGRAPHE SELON C. BLANCHE-BENVENISTE ET A. CHERVEL

C. Blanche-Benveniste et A. Chervel (1969), tout comme N. Catach, excluent de leur analyse les mots d'origine étrangère ainsi que les noms propres qui incluent des traits archaïques qui n'ont plus cours actuellement. Ils excluent également les phénomènes liés aux contraintes imposées par le matériau graphique. Par exemple, le phonème /ɛ/ n'est jamais noté *è* en finale de mot, principalement en raison de sa proximité graphique avec le graphème *é*. En lecture rapide, le lecteur risquerait de confondre des formes comme /tɔ̃be/ (*tombé*) et /tɔ̃bɛ/ (*tombait*) si cette dernière était graphiée **tombè*. En revanche, en milieu de mot, il n'y a souvent pas d'équivalent et donc pas de confusion possible, à l'exemple de /Rɛgl/ (*règle*) qui ne peut être confondu avec */Regl/ (**régle*).

Une fois cet aspect graphique écarté, leur ouvrage présente une orthographe à deux niveaux : le code phonographique et l'idéographie. Le code phonographique a vocation à exprimer la valeur phonique d'un graphème selon son contexte d'emploi. Il lie donc à chaque graphème un ensemble de phonèmes ou d'implications dans la chaine sonore. Rappelons que pour ces auteurs, le graphème est composé d'une lettre unique. L'idéographie est le principe inverse qui, pour une représentation phonique donnée, sélectionne une graphie parmi différents graphies possibles.

Le code phonographique

Le niveau du code phonographique permet de préciser la relation qui lie le graphème à la chaine sonore. Un graphème peut ainsi prendre cinq valeurs différentes :

– La **valeur de base** est « celle qui est liée au minimum de contraintes ». Ainsi le graphème *c* a pour valeur de base /k/, étant donné qu'il ne se réalise en /s/ que devant *i* et *e*.

– La **valeur de position** est une valeur conditionnée par la position du graphème. Le graphème *c* a pour valeur de position /s/ devant *i* et *e* (également *é*, *è* et *ê*). Un même graphème peut avoir plusieurs valeurs de position, ainsi *e* qui a pour valeur de base /ə/, peut prendre pour valeur de position les valeurs /e/, lorsqu'il est suivi d'un *r* ou d'un *t*, et /**ɛ**/ dans certains contextes, comme dans *mettre* ou *lemme*.

– La **valeur auxiliaire** : un graphème qui a une valeur auxiliaire n'a pas de correspondant phonique mais il influe sur la valeur d'un de ses voisins. Dans le mot *mangea*, *e* a une valeur auxiliaire : il ne se prononce pas mais conditionne la valeur de *g* qui prend sa valeur de position.

– **digramme** : dans un digramme deux graphèmes associés présentent une valeur phonique que ni l'un ni l'autre ne présente isolément. Par exemple dans la forme *ton*, *o* et *n* ont tous deux pour valeur /ɔ̃/, car à aucun moment le graphème *o* ou le graphème *n* ne peut prendre cette valeur en dehors de ce contexte *-on-*. Cette condition explique que, pour les auteurs, il n'existe pas de valeur trigramme. En effet, prenons le cas du groupe *ain* (/ɛ̃/). Le groupe *in* peut prendre la même valeur phonique /ɛ̃/ en l'absence du *a*. Le graphème *a* est donc considéré comme ayant une valeur zéro (cf. paragraphe suivant) ou une valeur auxiliaire selon les contextes (exemple *pain* ou *gain*). De même, le groupe *er* final n'est pas considéré comme un digramme puisque le graphème *e* a même valeur lorsqu'il précède un *t* final. Dans ces groupes, *e* a une valeur de position, *t* et *r* ont une valeur auxiliaire.

– La **valeur zéro** n'a pas de conséquence dans la relation graphie-phonie, elle correspond à des graphèmes qui ne se prononcent pas et qui n'influent pas sur la prononciation d'autres graphèmes.

La valeur que prend un graphème est donc modifiée par sa position dans le mot, cependant tous les graphèmes ne présentent pas le même nombre de valeurs. Ainsi, les graphèmes *j* et *v* ont une unique valeur, la valeur de base /ʒ/ et /v/, tandis que le graphème *o* présente 8 valeurs distinctes. C. Blanche-Benveniste et A. Chervel (1969) ont résumé les valeurs de chaque graphème dans le tableau suivant :

	Base	ph.	Position	ph.	Auxiliaire	Digramme	Zéro
A	*a*rt	a			améric*a*in g*a*in	au /o/, ai /e/, /ɛ/, an am /ɑ̃/, ay /ei/	p*a*in
B	*b*ar	b					plom*b*
C	*c*ar	k	*c*ire	s	exc*i*ter	ch /š/	banc
D	*d*ur	d			pie*d*		fon*d*
E	b*e*lette	ə	compl*e*t mang*e*r	ɛ e	gris*e* douc*e*âtre, g*e*ai, étai*e*nt	eu /ø/, ei /e/ en em /ɑ̃/, œ /e/ (ey)	boul*e*vard sol*e*, b*e*au
F	*f*er	f			cl*e*f		bœu*f*s
G	gare	g	*g*el	ž		gn /ny/	poing, vingt
H	*h*ibou	h			c*h*iromancie g*h*etto, éba*h*i	ch /š/, ph /f/ (sh)	*h*omme
I	*î*le	i	p*i*ed	y		ai /e, ɛ/, in im /ɛ̃/, il ill /y/, ei /e/, oi /wa/	o*i*gnon
J	*j*oli	ž					
K	*k*épi	k					stoc*k*
L	*l*it	l				ll ill il /y/	fi*l*s
M	*m*ère	m				am em /ɑ̃/, im ym /ɛ̃/ om /õ/ um /œ̃/	auto*m*ne da*m*ner
N	*n*u	n				an en /ɑ̃/ in yn /ɛ̃/ on /õ/ un /œ̃/ gn /ny/	ma*n*ne
O	*o*r	o	p*o*êle	u	c*œ*ur	œ /e/, oi /wa/, on om /õ/, ou /u/	ta*o*n
P	*p*ort	p				ph /f/	cham*p*
Q	*q*uand	k					cin*q*
R	*r*oi	r			aime*r*		ga*r*s, beu*r*re
S	*s*age	s	va*s*e	z	le*s*	(sh)	jeune*s*
T	*t*are	t	ac*t*ion	s	compl*e**t*		por*t*
U	*u*sine	ü	aq*u*atique	u	c*u*eillir g*u*êpier	au /o/ eu /ø/ ou /u/ un um /œ̃/	fatig*u*ant
V	*v*ase	v					
W	*w*agon	v					
X	a*x*e	ks	e*x*emple si*x* deu*x*ième	gz s z			deu*x*
Y	l*y*s	i	coba*y*e	y		yn ym /ɛ̃/, ay /ei/, (ey)	
Z	*z*èbre	z			ne*z*		ra*z*

Tableau 11. Les valeurs des graphèmes, C. Blanche-Benveniste et A. Chervel (1978, p.134)

L'idéographie

Selon C. Blanche-Benveniste et A. Chervel (1969), l'aspect idéographique de ce modèle permet de lier les différentes graphies d'un même segment phonique à un réseau de mots dérivés, ce qui permet de réaliser un choix orthographique, c'est-à-dire de différencier ces graphies en contexte. Ce choix se fait sur des critères dits « idéographiques » comme la dérivation au sein d'une famille de mots. Les auteurs donnent pour exemple la chaine sonore /pɛ̃/ qui peut prendre de nombreuses graphies parmi lesquelles *pin*, *pain* et *peint*. Le graphème *a* présent dans *pain* et qui a ici une valeur zéro permet de relier cette graphie aux dérivés *panier* et *panade* avec lesquels ils partagent trois graphèmes en commun : *p*, *a* et *n*. Si l'« idéographie » influe sur le code phonographique, l'inverse est également vrai puisque c'est

le code phonographique qui permet l'ajout du graphème *t* dans *vert* pour le lier à *verte*, mais ne permet pas l'ajout d'un graphème *m* à *ver* pour le lier à *vermisseau*. Ces deux niveaux sont donc complémentaires.

Ce modèle est principalement basé sur des principes phonographiques et discute très peu de l'aspect sémiographique, réduit à un principe idéographique. Pour une revue plus détaillée de ces phénomènes, nous nous tournons vers le modèle développé par N. Catach (1978) au sein de l'équipe HESO.

4.2.2.2. LE PLURISYSTÈME SELON N. CATACH ET L'ÉQUIPE HESO

Le plurisystème est un ensemble de systèmes inter-reliés qui permet de décrire les graphèmes selon le schéma suivant :

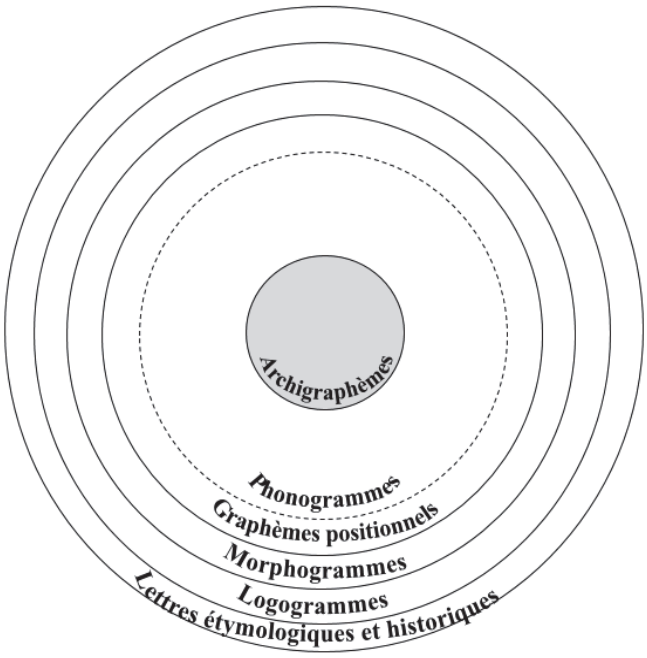


Figure 4. Le plurisystème, N. Catach (1978, p.55)

Les phonogrammes

La catégorie la plus centrale du schéma correspond à la catégorie des phonogrammes. Celle-ci inclut « tout graphème susceptible d'avoir un correspondant phonique » (Catach, 1979, p.27). Notre écriture est une écriture à principe alphabétique, les signes graphiques ont pour fonction première de transcrire les sons de la chaine parlée. La catégorie des phonogrammes est donc la plus dense, incluant 83,50% des graphèmes (Catach, 1978).

Le noyau de cette catégorie est constitué des archigraphèmes. Un ***archigraphème*** est un graphème fondamental choisi pour représenter un ensemble de graphèmes transcrivant le

même phonème (ou archiphonème). Le nombre d'archigraphèmes est donc lié au système phonologique de la langue.

Il est généralement admis que la langue française standard emploie 17 consonnes, 16 voyelles et 3 semi-consonnes, ou semi-voyelles (Fayol et Jaffré, 2008). Cependant, plusieurs raisons amènent N. Catach (1995), à réduire ce chiffre. En effet, de même que la langue évolue au cours du temps, sa prononciation évolue également. Certaines oppositions, de longueur et d'ouverture, ne se font plus ou uniquement dans quelques rares paires minimales. Ainsi d'aucuns distinguent *saule* de *sol*, tandis que d'autres opposent *patte* à *pâte*, mais cette opposition n'est plus réalisée que par une minorité des locuteurs. Les différences de prononciation dépendent souvent de la région d'origine et de l'âge du locuteur.

Dans le cadre de notre corpus, élaboré à échelle nationale, les scripteurs proviennent de différentes régions de France ou d'ailleurs et présentent sans doute de nombreux accents différents, il est donc difficile de faire un choix parmi les oppositions à retenir. D'où la question que posent C. Blanche-Benveniste et A. Chervel (1969, p.123) : « Faut-il tenir compte du maximum d'oppositions ou du minimum ? »

En se penchant sur les productions, il devient évident que de nombreux enfants scripteurs ne tiennent pas compte de ces différences. Et pour cause, les confusions sont fréquentes entre les graphèmes *é* et *ai* finaux, à l'exemple de « fé » (**680**, **1184**, **1280**, **1531**, …, *fait*). Cependant, sans analyse plus approfondie, il est difficile de savoir s'il s'agit d'une absence de différenciation entre ces phonèmes à l'oral ou une confusion entre les graphèmes à l'écrit.

N. Catach propose un système phonologique minimum composé de 17 consonnes, auxquelles on peut ajouter le phonème /ŋ/ présent dans *parking*, 3 semi-voyelles et 11 voyelles dont 3 sont accompagnées de variantes de position. Si d'ordinaire les sons de la langue sont représentés par des phonèmes, ces dernières voyelles sont représentées par des archiphonèmes. Un archiphonème est un « représentant de l'ensemble des traits phoniques pertinents communs à deux ou plusieurs phonèmes, qui sont par rapport aux autres dans un rapport exclusif » (Catach, 1995, p.16). A sa suite, nous considèrerons un système phonologique restreint (tableau 12). Mais nous ne considèrerons que 10 voyelles, la voyelle /œ/ étant pour nous une variante positionnelle de la voyelle /ɛ̃/.

Voyelles		Consonnes		Semi-voyelles
Voyelles orales	**Voyelles nasales**	/p/	/b/	
/a/	/ɑ̃/	/t/	/d/	
/E/ : /e/ + variante positionnelle /ɛ/		/c/	/g/	
/i/		/f/	/v/	/j/
/O/ : /o/ + variante positionnelle /ɔ/	/ɔ̃/	/s/	/z/	
/y/		/ʃ/	/ʒ/	/ɥ/
/œ/ : /ø/ + variantes positionnelles /œ/ et /ə/	**/œ̃/ : /ɛ̃/ + variante positionnelle /œ̃/**	/l/	/r/	
/u/		/m/	/n/	/w/
		/ɲ/	(/ŋ/)	

Tableau 12. Système phonologique restreint du français

Ce système est un système à 30 ou 31 phonèmes pour lesquels nous disposons de 26 lettres simples, de 5 signes diacritiques (les 3 accents, les trémas, la cédille) répartis sur 13 lettres (ç, é, à, è, ù, ë, ï, ü, â, ê, î, ô, û) soit 39 signes (Blanche-Benveniste et Chervel, 1969). Ce chiffre varie selon les auteurs, N. Catach (1995) compte 43 signes (en incluant œ, ÿ, ñ et *æ* dont nous ne tiendrons pas compte). Cependant, C. Blanche-Benveniste et A. Chervel nous mettent en garde sur le fait que certains de ces signes ne sont pas des graphèmes à part entière et qu'ils n'ont qu'une fonction « idéographique » à l'image du *ù* qui permet de distinguer *ou* et *où.*

Les morphogrammes

Vient ensuite le groupe des morphogrammes. Ces graphèmes ne correspondent plus à des phonèmes mais à des morphèmes, que ce soit de désinence, de flexion, de dérivation, d'affixes, etc. Ces graphèmes peuvent être prononcés ou non et dépendent étroitement des liaisons et relations établies en contexte. En lexique, cette catégorie ne représente que 3,67% des graphèmes, mais cette proportion augmente dans les textes.

N. Catach distingue les morphogrammes grammaticaux des morphogrammes lexicaux. Les morphogrammes grammaticaux sont des marques graphiques de flexion ou de désinence qui peuvent s'ajouter au mot lorsque celui est placé en discours. Les morphogrammes lexicaux sont des marques graphiques intégrées au lemme, ils peuvent marquer les dérivations et les flexions, à l'exemple du *d* de $\text{RENARD}_{\text{NOM}}$ ou du *t* dans $\text{SOURIANT}_{\text{ADJ}}$ qui permettent de faire un lien avec les formes féminin *renarde* et *souriante* (Catach, 1978), ou correspondre à des marques d'affixes comme le graphème *im* de $\text{IMMANGEABLE}_{\text{ADJ}}$.

Les logogrammes

Vient alors une zone plus annexe contenant les graphèmes appelés logogrammes ou « figure de mots ». Les logogrammes sont généralement des unités constituées d'un ou deux

phonèmes (monosyllabe) ou de mots très fréquents. Ils permettent de distinguer certains homophones à l'aide de graphies particulières. Par exemple le signe diacritique de *où* permet de distinguer cette unité de l'unité lexicale *ou*. Ces graphèmes sont peu fréquents et ne concernent que 3,27% des mots.

Lettres étymologiques et historiques

Enfin, il existe un petit nombre de lettres utilisées mais qui ne répondent pas à tous les critères donnés dans la définition du graphème. Ces lettres seront qualifiées de « lettres étymologiques et historiques ». On retrouve dans ce cas de nombreuses consonnes doubles qui ne s'expliquent pas par des phénomènes de construction linguistique comme l'affixation. On donnera pour exemple les mots *sonner*, *donner*, *honneur*, etc. (Catach, 1995) pour lesquels le doublement du *n* ne peut s'expliquer que par la connaissance de l'étymologie ou de l'histoire.

Pour conclure, ces deux modèles décrivent tous les deux le même objet, à savoir l'orthographe. Cependant, ils ne le font pas à partir de la même unité d'étude, ni à partir de la même approche. En effet, la théorie développée par C. Blanche-Benveniste et A. Chervel s'appuie sur une définition graphique du graphème pour en faire une description basée sur le principe phonographique. La théorie proposée par N. Catach, quant à elle, utilise une définition basée à la fois sur la phonographie et la sémiographie pour en étudier les fonctions (tableau 13). L'orthographe est donc un système complexe que les scripteurs doivent apprendre à maitriser, c'est sur cet apprentissage que nous nous pencherons dans le chapitre suivant.

	Blanche-Benvéniste et Chervel	Catach (HESO)
Phonographie	- valeur de base - valeur de position - valeur auxiliaire - valeur digramme - valeur zéro	Phonogrammes
Sémiographie	Idéographie	- Morphogrammes - Logogrammes - Lettres étymologiques et historiques

Tableau 13. Comparaison des théories orthographiques de C. Blanche-Benveniste et N. Catach

CHAPITRE 5 - L'APPRENTISSAGE DE L'ÉCRITURE

Après nous être intéressés au système d'écriture du français, nous pouvons désormais étudier la façon dont les enfants entrent dans le monde de l'écrit, s'approprient le système graphique de la langue et apprennent à écrire selon la norme orthographique.

5.1. ACQUISITION DU SYSTÈME D'ÉCRITURE

Dès l'âge de trois ans, les enfants produisent des formes graphiques qu'ils distinguent du dessin. De nombreux modèles ont été avancés pour représenter l'évolution entre ces premières traces d'écrits et l'écrit tel que nous l'utilisons selon la norme du français (cf. Geoffre, 2013, p. 64, pour une revue et une synthèse de ces modèles). La plupart de ces modèles présentent une acquisition par stades ou par étapes et suggéreraient l'idée selon laquelle l'enfant devrait passer successivement par chacune des étapes. Or, cette vision étapiste est de plus en plus remise en question, un enfant pouvant utiliser plusieurs procédures en synchronie. Si nous ne pouvons développer ici chacun de ces modèles, nous pouvons tout de même développer le modèle proposé par U. Frith (1985) qui a influencé la façon dont on a observé l'acquisition de l'écriture en langue française.

5.1.1. LE MODÈLE D'UTA FRITH

U. Frith (1985) a développé un modèle en trois phases qui tente de montrer l'interdépendance qui lie lecture et écriture. Détaillons ces trois phases :

– Le **stade logographique** : lors de ce stade, l'enfant reconnait et mémorise les mots qui lui sont présentés à l'aide d'indices visuels saillants (lettre reconnue, configuration, etc.) ou contextuels (couleur, position, etc.). Il ne s'agit pas encore de lecture mais d'une reconnaissance « globale » ou « visuelle » des mots. Cette procédure permet de reconnaitre certains mots apparaissant toujours sous le même aspect (comme les logos publicitaires par exemple) mais aussi certains mots familiers. Ce stade permettrait de constituer un premier lexique de 10 à 100 mots (Chanquoy et Negro, 2004).

– Le **stade alphabétique** : les correspondances phonographiques commencent à émerger. Il y a donc une prise de conscience à la fois des unités de l'écrit et des unités de l'oral et du lien qui les lie. L'enfant reconnait alors de plus en plus de mots, son lexique mental peut alors croitre.

–Le **stade orthographique** : l'enfant prend conscience des limites de la stratégie alphabétique. Il emmagasine de nouveaux mots en mémoire à long terme, notamment les mots irréguliers. Cette étape lui permet ensuite de reconnaitre ces mots par voie directe, sans passer par un décodage phonographique. Contrairement au stade logographique, l'adressage se fait par traitement sémantique et non plus par traitement visuel. La mémorisation des mots nouveaux est facilitée par un procédé analogique qui permet de les comparer, au moins en partie, à des mots déjà connus. À ce stade, c'est la lecture qui permet à l'enfant d'accroitre son lexique mental et ses connaissances orthographiques. Plus ce stade est automatisé, plus il inclut de mots et plus l'enfant est considéré comme expert à l'écrit.

Ce modèle a depuis été critiqué à de nombreuses reprises, notamment sur son aspect en stade. C. Martinet, S. Valdois et M. Fayol (2004, cités dans Pacton, Foulin et Fayol, 2005) ont réalisé des travaux dans le but de montrer l'indépendance des deux derniers stades, notamment en montrant que la maitrise des correspondances phonographiques n'est pas un prérequis obligatoire à la procédure orthographique. En effet, ils ont montré, à partir d'une expérience réalisée après trois mois et après neuf mois d'apprentissage, que les mots irréguliers les plus fréquents étaient très tôt mieux orthographiés que les autres mots irréguliers. Ils ont également montré que les enfants, surtout ceux possédant de bonnes connaissance lexicales, sont déjà capables d'écrire en ayant recours à des phénomènes d'analogie. Ces observations tendent à montrer une mémorisation de traces orthographiques, avant même la maitrise totale du principe alphabétique.

5.1.2. ÉCRITURES INVENTÉES

J. Fijalkow (2009, p.63) définit les écritures inventées comme étant « un graphisme à partir de ce qu'ils [les enfants] pensent être l'écriture et à l'aide des connaissances dont ils disposent ». Le concept d'*écritures inventées* ou *orthographes inventées* (« invented spelling ») a été initié en langue anglaise par C. Chomsky (1971) et C. Read (1971). E. Ferreiro (Ferreiro et Teberovsky, 1979, Ferreiro et Gomez-Palacio, 1988, Ferreiro, 2000), reprenant et développant leur travaux, s'est intéressée à la façon dont les enfants conçoivent et s'approprient l'écrit et à l'évolution de leurs graphies.

Les travaux menés par E. Ferreiro ont permis de développer un modèle centré principalement sur la façon dont l'enfant acquiert la graphie avant le stade alphabétique (cf. modèle d'U. Frith) et qui présuppose que l'enfant traverse différents stades :

– Le **stade pré-syllabique** : stade où l'enfant apprend à différencier l'écriture des marques iconiques, il comprend que l'écrit n'est pas du dessin. Pour l'enfant, l'écriture a une fonction de référence par rapport à une entité donnée. Il va alors attribuer à ses écrits certaines caractéristiques des entités à représenter, à l'exemple de la longueur de la production selon la taille de l'animal. Les productions graphiques ressemblent alors à des gribouillis ou des lignes en formes de vagues, avant de présenter des pseudo-lettres ou les lettres les mieux connues de l'enfant, c'est-à-dire celles de son prénom. Durant ce stade, l'enfant comprend également que pour représenter des entités différentes, les représentations graphiques doivent être différentes (Ferreiro, 1988).

– Le **stade syllabique** : l'enfant comprend que ce sont les aspects sonores qui sont transcrits par la graphie. La chaine sonore est découpée en syllabes et chaque syllabe est transcrite par une lettre qui n'a pas forcément de lien avec la syllabe elle-même. À l'exemple de Jorge, 6 ans, qui écrit **AEI* pour *Ca-ba-llo* (*cheval*), ou encore **EI* pour *Ga-to* (*chat*). Du point de vue du français, C. Bégin, L. Saint-Laurent et J. Giasson (2005) nous donnent ainsi l'exemple de *bâ-to-nnet* écrit **eio*.

– Le **stade syllabico-alphabétique** : La lettre utilisée pour transcrire une syllabe est maintenant contenue dans la syllabe, à l'exemple de Martin, 6 ans, qui écrit **aioa* pour *Ma-ri-po-sa* (*papillon*). L'enfant ajoute progressivement des graphèmes en plus, désormais une syllabe peut être représentée par plusieurs lettres. Il s'agit d'un stade intermédiaire où certaines syllabes sont partiellement transcrites tandis que tous les phonèmes des autres sont transcrits, par exemple **lais* pour *lá-piz* (*crayon*). L'exemple précédent *bâtonnet* peut alors être écrit **batn* (Bégin *et al.*, 2005).

– Le stade **alphabétique** : le principe alphabétique est acquis, l'enfant transcrit tous les phonèmes par des graphèmes, même si cette transcription ne respecte pas encore la norme orthographique, à l'exemple de la production **éléfen* (éléphant) (Bégin *et al.*, 2005).

Depuis lors, l'hypothèse qu'émet E. Ferreiro (1988) et qui présuppose un découpage en syllabes de la part de l'enfant a été remise en question, notamment pour des langues où le nombre de syllabes est bien supérieur à celui de l'espagnol, avec une structure syllabique beaucoup plus diversifiée, comme le français.

Ces travaux ont ensuite été repris et adaptés au milieu scolaire afin d'étudier l'évolution des écritures sur un plan plus didactique. Ces travaux ont pris le nom d'*orthographes approchées* du fait d'une comparaison plus systématique à la norme

orthographique. J. David et S. Fraquet (2011) en donnent la définition suivante : « pratiques scolaires permettant aux élèves de produire directement et de manière autonome des écrits [...]. Au-delà de ces diverses appellations, l'objectif central est d'amener les jeunes élèves à produire des écrits par résolution progressive des problèmes linguistiques et (ortho)graphiques ».

Tous ces travaux montrent que très tôt les enfants présentent un intérêt pour l'écrit et une certaine compétence avant même un apprentissage formel. De plus, l'apprentissage de l'écrit ne réside pas seulement dans l'apprentissage du système de représentation que constitue l'écrit mais également dans la construction au niveau individuel de ce système de représentation (Ferreiro, 2008). En clair, il ne s'agit pas seulement pour l'enfant d'apprendre à encoder les données de la langue, mais il lui faut d'abord comprendre quelles sont les informations à représenter. C'est pourquoi avant de construire le lien entre la graphie et la phonie, l'enfant va, par exemple, faire des liens entre la graphie et l'entité physique représentée, comme le fait par exemple l'enfant qui attribue plus de lettres à *ours* qu'à *papillon* sous le prétexte que l'ours est un animal plus grand que le papillon (Bégin *et al*., 2005).

5.1.3. LES TRAITEMENTS DE L'ÉCRIT

J. Fijalkow et A. Liva (1993) qui s'intéressent également aux modalités d'entrée dans l'écrit ne parlent plus de stades mais de traitements, bien qu'ils se succèdent également dans un ordre chronologique. Ces traitements vont reprendre les grandes observations des deux modèles précédents. Ainsi, l'entrée dans le monde de l'écrit se fait par un traitement d'abord exclusivement visuel, il s'agit des traitements figuratif et visuel, vient ensuite un traitement incluant l'oral et enfin un traitement idéo-visuel, appelé traitement orthographique.

– Le **traitement figuratif** : l'enfant perçoit l'écrit comme une représentation de ce qui et dit à l'oral et dessine la forme entendue. Puis, il comprend la différence entre écriture et dessin et réalise un tracé qu'il apparente à de l'écriture.

– Le **traitement visuel** : l'enfant affine de plus en plus sa façon de se représenter l'écrit. Il fait d'abord une simulation de l'écrit à l'aide de pseudo-lettres puis a de plus en plus recours aux lettres conventionnelles, qui sont, dans un premier temps, majoritairement celles de son prénom.

– Le **traitement de l'oral** : l'enfant fait à la fois une analyse en phrases et en mots. Les phrases sont écrites avec plus de lettres que les mots. La phrase peut alors être écrite de

plusieurs façons : l'enfant peut écrire une lettre par mot ou segmenter la phrase en différentes parties constituées de mots ou groupes de mots. Progressivement, la phrase contient autant de parties que de mots. De même, l'identification du mot s'affine progressivement. Alors qu'il n'y a au début qu'une lettre par syllabe, des correspondances grapho-phonétiques apparaissent peu à peu et notamment à l'attaque des mots, jusqu'à l'apparition d'une écriture phonétique. Une écriture est dite phonétique lorsqu'apparait au moins une lettre par phonème. Au début, seuls certains mots parmi les plus courts sont écrits phonétiquement, puis progressivement cette proportion s'accroit. Au cours de cette phase, la perception de l'oral s'affine, l'enfant peut désormais identifier des syllabes et des phonèmes.

– Le **traitement orthographique** : l'enfant comprend les limites du traitement oral et bien que ce traitement reste majoritaire, il tient également compte d'indices visuels relevant de l'orthographe. D'abord partiel (quelques mots), ce traitement est de plus en plus systématique (phrase entière), bien que quelques difficultés puissent persister jusqu'à un âge plus avancé, notamment entre graphèmes homophones ou sur les morphogrammes grammaticaux (Fijalkow, Cussac-Pomel et Hannouz, 2009).

Afin de mieux situer les différentes approches les unes par rapport aux autres, nous pouvons en proposer un résumé à l'aide du schéma suivant :

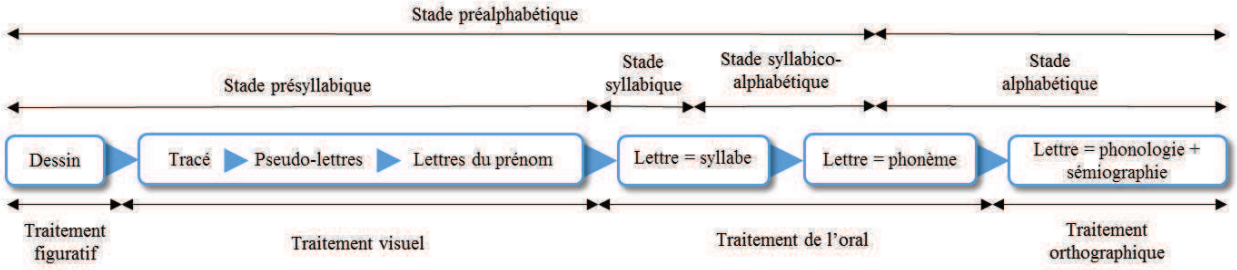


Figure 5. Résumé des modèles approchant les écritures inventées

Les productions présentes dans notre corpus ne peuvent plus être considérées comme des écritures inventées ou orthographes approchées puisqu'elles ont été recueillies après un an d'apprentissage institutionnel de l'écriture. Toutefois, on retrouve encore dans notre corpus certaines procédures employées dans les orthographes approchées, bien que, comme nous le verrons par la suite, la procédure majoritaire consiste à passer par des correspondances phonie-graphie. Étudier les orthographes approchées peut donc aider à la compréhension de notre corpus. Les productions présentant des caractéristiques des phases présyllabiques et syllabiques (3 productions, cf. annexe 6) ont été annotées lors de la transcription au moyen de la balise <letDET/> mais ne peuvent être traitées automatiquement, nous les laisserons donc de côté à l'avenir, de même que les 12 productions vides ou ne contenant qu'un dessin ou le prénom. Les productions correspondant au stade syllabico-alphabétique (10 productions) tiendront certainement un système d'annotation

automatique en échec. Bien qu'elles aient été transcrites, nous nous concentrerons donc principalement sur les productions aux stades alphabétique et orthographique.

5.2. LES DIFFICULTÉS EN DÉBUT D'APPRENTISSAGE

Certains mots utilisés par les élèves dans notre corpus sont des mots vus et mémorisés en classe de CP et pour lesquels un traitement orthographique est d'ores et déjà possible, par une référence directe en lexique mental. La plupart, en revanche, sont des mots dont l'orthographe leur est inconnue. Pour S. Fraquet et J. David (2013), l'écriture confronte les enfants à de nombreuses difficultés qui sont principalement la découverte et l'application du principe alphabétique de notre écriture et la segmentation en mots graphiques.

5.2.1. COMPRENDRE LE LIEN ENTRE ORAL ET ÉCRIT ET SES LIMITES

Une partie du travail du scripteur revient à encoder les sons de la chaine parlée, cela nécessite de se familiariser avec le principe alphabétique, c'est-à-dire à la transcription de syllabes et de phonèmes. Avant que le code phonographique soit entièrement acquis, différentes stratégies d'encodage sont employées par les enfants. J. David (2003) nous donne ainsi l'exemple d'une enfant écrivant **rdard* pour *regarder*. Dans cet exemple, la première syllabe écrite au moyen de la lettre *r* correspond à une notation syllabique où une lettre est utilisée pour transcrire la totalité de la syllabe. La deuxième syllabe, écrite *dar*, résulte d'une transcription par phonèmes, ici, chaque phonème est transcrit par une lettre. Pour finir, la dernière syllabe, écrite *d*, résulte d'une procédure épellative où le nom de la lettre est pris pour valeur.

Comme J.-P. Jaffré (Fayol et Jaffré, 2008) l'a écrit, une des difficultés auxquelles sont confrontés les scripteurs de langue française réside dans la grande polyvalence orthographique des signes utilisés et la complexité des correspondances phonographiques qui le caractérise. En 1978, Catach résume la multiplicité des correspondances possibles dans un tableau mettant en correspondance chaque archigraphème et ses graphèmes associés ainsi que leurs fréquences (annexe 7). Ainsi, le phonème /o/ est représenté par l'archigraphème *O* hors-contexte, tandis qu'en contexte on le trouve orthographié de plusieurs manières comme *o*, *au* et *eau*. Ce tableau nous montre que certaines correspondances phonie-graphie sont plus faciles à établir que d'autres. En effet, seul le graphème *u* permet de transcrire le phonème /u/ (archigraphème *U*), il n'y a donc aucune hésitation à avoir pour orthographier ce phonème, pour un locuteur natif du français ayant une perception des sons non déviante.

En revanche, comme nous venons de l'évoquer, trois graphies différentes permettent de transcrire le phonème [o], c'est donc une source d'erreurs potentielles. Outre les phonèmes sources d'erreurs, ce tableau nous donne également le pourcentage d'utilisation du graphème, selon le phonème à transcrire. Ainsi, selon Catach, pour transcrire le phonème /o/ en langue française, le graphème *o* est utilisé dans 75% des cas, tandis que les graphèmes *au* et *eau* ne sont utilisés qu'à hauteur de 21% et de 3%.

Pour faire un choix entre ces graphèmes, il est nécessaire de prendre en compte d'autres aspects de la langue que simplement le lien entre chaine sonore et chaine graphique. De même, le principe alphabétique ne permet pas d'expliquer la présence de lettres qui n'ont aucun équivalent dans la chaine sonore. Les jeunes scripteurs doivent également apprendre à prendre en compte des contraintes morphographiques, logographiques ou étymologiques.

Une étude menée par S. Pacton, M. Fayol et P. Perruchet en 2002 a montré que les enfants n'utilisant que deux de ces graphèmes pour transcrire le phonème /o/ utilisent les graphèmes *o* et *au*, tandis qu'un enfant n'utilisant qu'un seul de ces graphèmes utilise le graphème *o*. Mais cette étude a surtout montré que la fréquence de ces différents graphèmes est modifiée en fonction du contexte et que les enfants sont très tôt sensibles à ces différences d'utilisation en contexte. Dans cette expérience, il était demandé à des enfants de CE1, CE2 et CM1 d'écrire des pseudo-mots tri-syllabiques. Il en est ressorti, par exemple, que les enfants utilisaient davantage le graphème *eau* en position finale qu'en position médium ou initiale et plus souvent après un *v* qu'après un *f*. Or, ce sont des caractéristiques qu'on retrouve en français, il y a donc une prise en compte très tôt du contexte graphémique.

Sur la base des connaissances contextuelles et orthographiques qu'il est nécessaire d'acquérir pour écrire un mot, M. Fayol (2013) a établi une échelle de difficulté. Ainsi, les mots les plus faciles sont les mots dits « réguliers », c'est-à-dire qu'ils s'écrivent comme ils se prononcent, comme *ton* ou *tabou*. Viennent ensuite les mots qui contiennent des lettres muettes motivées morphologiquement à l'exemple de *renard*, dont le *d* le lie à la forme féminine *renarde*. Enfin, les mots considérés comme les plus irréguliers sont ceux qui contiennent des lettres muettes non motivées à l'exemple de *temps* ou encore les mots qui ne contiennent pas de lettres muettes mais qui pourraient en contenir au vu de leur morphologie, comme *abri*(t) qui, en lien avec sa forme dérivée *abriter*, pourrait contenir un *t* final.

5.2.2. APPROCHER LA SEGMENTATION

5.2.2.1. SEGMENTER UN TEXTE EN MOT

Une des difficultés majeures de l'écrit est la segmentation de la chaine parlée en mots distincts. C. Fabre-Cols (2000) notent que cette étape est rendue plus difficile par le fait qu'aucun indicateur phonique ou prosodique ne marque l'unité mot. En effet, les accentuations marquent des groupes phoniques et non des mots. De plus, il est important de noter que certains éléments compliquent encore la tâche de segmentation comme les liaisons, par exemple *un arbitre* prononcé comme *un narbitre* et les élisions, comme dans *l'ours*.

Différentes stratégies de segmentation peuvent être adoptées au cours de l'apprentissage (Fraquet et David, 2013). On observe notamment une stratégie de copie du mot, cette stratégie donne généralement lieu à une segmentation réussie puisque le mot est copié comme une unité isolée des autres mots. Une autre stratégie possible est le découpage phonographique qui donne lieu à une segmentation en syllabes. Enfin, une dernière stratégie consiste à segmenter l'écrit en groupes syntaxiques et ou sémantiques.

P. Cappeau et M.-N. Roubaud (2005) ont étudié les erreurs de segmentation et en ont observé quelques tendances générales. Ils ont montré qu'il y avait souvent agglutination d'un mot de la classe des « petits mots fréquents » avec un autre, adjacent. Ces petits mots peuvent correspondre à des déterminants, des prépositions, des conjonctions, des adverbes ou à des pronoms, par exemple **de lautre coter* (CE1), ou encore **la sorcière dit buvésa* (CE1). Mais les mots peuvent également être écrits en plusieurs segments différents. Ils observent que le découpage réalisé correspond généralement à des mots existants (tout du moins phonétiquement parlant), à l'exemple de **en dormi* (CE1) ou encore **le loup été pré ala trapé* (CP).

5.2.2.2. SEGMENTER UN TEXTE EN PHRASES

Les problèmes que pose la segmentation d'un texte ne se limitent pas à la segmentation en mots mais également en ***phrase***. P. Cappeau et M.-N. Roubaud (2005) distinguent la phrase graphique, celle qui est marquée par la ponctuation et les majuscules, de la phrase syntaxique, qui correspond à une structure syntaxique. Même si cette structure syntaxique peut être correctement réalisée, il est fréquent qu'aucun marqueur ne l'identifie comme une phrase graphique. À l'inverse, certaines productions d'élèves présentent une ou plusieurs phrases graphiques sans que celle-ci soit apparentée à une phrase syntaxique mais plutôt à une ligne ou au texte dans sa globalité. Ainsi, la segmentation en phrases graphiques à l'aide de

marqueurs tels que la ponctuation est encore très peu maitrisée dans les premières années d'école primaire.

D'après Passerault (1991), l'emploi du point commence à apparaitre dès le début de la maitrise de l'écrit qu'il situe au CE1, alors que la virgule n'apparait qu'au CE2. Cette dernière affirmation semble se confirmer dans notre corpus, qui ne comporte que très peu de virgules. Néanmoins, nous observons l'apparition du point chez certains élèves dès la classe de CP. Nous pouvons dès lors nous concentrer sur l'usage du point.

Bien que l'emploi du point par des scripteurs débutants ne corresponde pas à l'emploi qu'en ferait un scripteur adulte, les études ont montré que la place du point a déjà un caractère non aléatoire dès le début de l'apprentissage (Passerault, 1991). Ainsi, selon B. Schneuwly (1984, cité dans Passerault, 1991), l'usage du point a pour but de marquer une fin de phrase, toutefois, l'inverse n'est pas vrai : la fin d'une phrase n'est pas toujours indiquée par un point, voire peu souvent. Le caractère non systématique de l'utilisation du point en fin de phrase semble indiquer que sa valeur n'est pas toujours phrastique. En effet, P. Cappeau et M.-N. Roubaud (2005) montrent à travers leur corpus que le point est souvent attesté en fin de texte ou employé comme séparateur de deux épisodes narratifs. Un tel constat nous laisse supposer que le point et la ponctuation n'ont alors pas seulement une fonction phrastique mais surtout une fonction textuelle. Fayol (1989) avance ainsi que la première fonction de la ponctuation serait d'indiquer le degré de relation entre les éléments du texte.

Notons également que de nombreux auteurs (Lurçat, 1985) (Schneuwly, Rosat et Dolz, 1989, Fayol, 1981, 1986, cités dans Cappeau et Roubaud, 2005) mentionnent un emploi similaire entre connecteurs, principalement le connecteur *et* (Schneuwly *et al.*, 1989), et ponctuation chez les scripteurs débutants.

5.3. GRILLES ET TYPOLOGIES D'ERREUR

Afin de détecter les erreurs produites par les enfants lors de l'apprentissage, il est nécessaire de développer des typologies d'erreurs qui permettent d'évaluer les erreurs selon des critères linguistiques ou psycholinguistiques. Du fait de son utilisation en didactique et en TAL, nous présenterons ici le modèle développé par l'équipe HESO dirigée par N. Catach (1980). Ce modèle s'appuie sur les catégories développées précédemment dans la présentation du plurisystème (cf. 4.2.2.2.). Puis, nous présenterons la terminologie employée par P. Cappeau et M.-N. Roubaud (2005). Pour finir, nous nous pencherons sur les propositions de P. Guimard (2003).

5.3.1. TYPOLOGIE DE CATACH, DUPREZ ET LEGRIS

5.3.1.1. ERREURS À DOMINANTE PHONÉTIQUE

Qu'elles soient dues à des problèmes d'audition ou des problèmes de prononciation, les erreurs à dominante phonétique modifient la valeur phonique des mots qui les portent. Sont classées parmi ces erreurs les omissions ou adjonctions de lettres ou de syllabes, ainsi que les confusions entre phonèmes. Contrairement aux erreurs qui vont suivre, ces erreurs ne sont pas qualifiées d'erreurs graphiques mais d'erreurs extragraphiques.

Catégories d'erreurs	Remarques	Exemples
Erreurs extragraphiques		
0. Erreurs à dominante calligraphique	Ajout ou absence de jambages, etc.	* mid/ (nid)
0 *bis.* Reconnaissance et coupure des mots	Peut se retrouver dans toutes les catégories suivantes	Le *lévier (l'évier)
1. ERREURS A DOMINANTE EXTRAGRAPHIQUE (en particulier phonétique) - enrichir la grille des principales oppositions des phonèmes (voyelles, semi-voyelles ; consonnes)	- Omission ou adjonction de phonèmes - Confusion de consonnes - Confusion de voyelles : ex. [ɔ]/ [ə]	* maitenant (maintenant) * suchoter (ch/s) * moner (mener)

Tableau 14. Extrait 1 de la grille typologique des erreurs d'orthographe (N. Catach, 1995, p. 282)

5.3.1.2. ERREURS À DOMINANTE PHONOGRAMMIQUE

Ces erreurs sont classées en deux catégories selon qu'elles altèrent la valeur phonique ou non. À l'intérieur de ces catégories, les erreurs sont distinguées selon la nature sonore du phonème sur lequel elles portent (voyelle, semi-voyelle, consonne ainsi que les consonnes doubles ou simples pouvant être doublées qui sont classées à part des consonnes) et selon la nature de l'erreur (omission ou adjonction, confusion ou inversion). Toutes ces catégories permettent de donner une description très fine de l'erreur mais complexifie le travail de classement (Catach, 1980), de nombreuses erreurs pouvant être placées dans plusieurs catégories à la fois.

Catégories d'erreurs	Remarques	Exemples
Erreurs graphiques proprement dites		
2. ERREURS A DOMINANTE PHONOGRAMMIQUE (règles fondamentales de transcription et de position) - enrichir la grille en se fondant sur les archigraphèmes (voyelles, semi-voyelles ; consonnes)	- Altérant la valeur phonique - N'altérant pas la valeur phonique	* merite (mérite) * briler (briller) * recu (reçu) * binètte (binette) * pingoin (pingouin) * guorille (gorille)

Tableau 15. Extrait 2 de la grille typologique des erreurs d'orthographe (N. Catach, 1995, p. 282)

5.3.1.3. ERREURS À DOMINANTE MORPHOGRAMMIQUE

Ces erreurs sont également décrites de manière très fine. En premier lieu, les morphèmes grammaticaux sont distingués des morphèmes lexicaux. Les morphèmes grammaticaux sont à leur tour différenciés selon la catégorie et le voisinage syntaxique du lexique sur lequel ils portent. Les morphèmes lexicaux sont classés selon le degré de reconnaissance supposé et de la maitrise orthographique du mot (selon qu'ils portent les marques des affixes, de la famille lexicale, des lettres finales, etc.).

Catégories d'erreurs	Remarques	Exemples
Erreurs graphiques proprement dites		
3. ERREURS A DOMINANTE MORPHOGRAMMIQUE - enrichir la grille en se fondant sur les principaux morphogrammes et les principales catégories d'accord *1. Morphogrammes grammaticaux*	- Confusion de nature, de catégorie, de genre, de nombre, de forme verbale, etc. - Omission ou adjonction erronée d'accords étroits - Omission ou adjonction erronée d'accords larges	* chevaus (chevaux) * les rue (les rues) Ceux que les enfants ont *vu (vus)
2. Morphogrammes lexicaux	- Marques du radical - Marques préfixes/suffixes	* canart (canard) * anterremant (enterrement) * annui (ennui)

Tableau 16. Extrait 3 de la grille typologique des erreurs d'orthographe (N. Catach, 1995, p. 282)

5.3.1.4. ERREURS CONCERNANT LES HOMOPHONES, LES IDÉOGRAMMES ET LES LETTRES NON FONCTIONNELLES

Les erreurs d'homophonie concernent à la fois les homophones lexicaux, grammaticaux et les homophones en discours, c'est-à-dire des séquences de mots qui sont homophones sans avoir le même découpage en mots. Les erreurs concernant les idéogrammes portent principalement sur les majuscules, sur l'apostrophe et le trait d'union et enfin sur la ponctuation. Enfin, les lettres non fonctionnelles présentant des erreurs sont classées selon qu'elles soient des voyelles, des consonnes justifiées par l'étymologie ou des lettres injustifiées ou difficilement justifiables.

Catégories d'erreurs	Remarques	Exemples
Erreurs graphiques proprement dites		
4. ERREURS A DOMINANTE LOGOGRAMMIQUE	- Logogrammes lexicaux - Logogrammes grammaticaux	j'ai pris du *vain (vin) ils *ce sont dit (se)
5. ERREURS A DOMINANTE IDEOGRAMMIQUE	- Majuscules - Ponctuation - Apostrophe - Trait d'union	l'*état (l'État) * et, lui (et lui) * létat (l'État) * mot-composé (mot composé)
6. ERREURS A DOMINANTE NON FONCTIONNELLE	- lettres étymologiques - consonnes simples ou doubles non fonctionnelles - accent circonflexe (non distinctif)	* sculteur, *rume (sculpteur, /rhume) * boursouffler (boursoufler) * anerie, *patisserie

Tableau 17. Extrait 4 de la grille typologique des erreurs d'orthographe (N. Catach, 1995, p. 282)

5.3.2. UNE TYPOLOGIE À DEUX ÉTAGES

En rupture avec ces travaux et reprenant les travaux d'A. Chervel et de C. Blanche-Benveniste, P. Cappeau et M.-N. Roubaud (2005) proposent une différenciation à deux étages :

– Les erreurs de **code phonographique** portant sur la mise à l'écrit d'un mot oralisé. Ce niveau d'analyse implique deux questions, celle de savoir avec quelle graphie transcrire un son (une erreur à ce niveau-là impliquerait une erreur que l'on pourrait appeler « erreur de transcription phonographique ») et celle de savoir comment découper un mot (ici l'erreur serait une erreur de segmentation).

Les erreurs de sélection de la **norme orthographique**. Au niveau appelé « orthographe », il est nécessaire de sélectionner une graphie parmi plusieurs graphies possibles, selon la norme orthographique.

5.3.3. LA CLASSIFICATION DE P. GUIMARD (2003)

Du côté de la psychologie du développement, P. Guimard (2003) a proposé un modèle permettant d'établir des profils selon les erreurs produites. L'intérêt de ce modèle est que, contrairement aux modèles précédents, il ne classe pas les erreurs au niveau du graphème, mais au niveau du mot (dans son cas, une production étant égale à un mot). Or, dans le schéma que nous allons élaborer par la suite (cf. chapitre 6), il nous faudra à la fois répertorier l'erreur au niveau du graphème, mais également au niveau du segment. Il distingue ainsi six catégories de productions :

– les productions **correctes** ;

– les productions « **oralisables** » qui se décomposent en formes « pures » (**balon*) et en formes complexes lorsqu'il s'agit d'une transcription proche (**ten*, forme normée *dent*) ou plausible dans un autre contexte lexical (**glase*, forme normée *glace*). En effet, *s* permet également de transcrire le phonème /s/ dans un autre contexte lexical, comme *sapin* ;

– les productions « **reconnaissables** » lorsque la forme phonologique n'est pas la bonne (**tambou*, forme normée *tambour*) ;

– les productions **autres**, dont la forme est très éloignée de la forme normée (**p*, forme normée *poire*) ;

– les **non productions** ;

– les **variations lexicales** (*peigne* remplaçant *brosse*).

5.3.4. CONCLUSION

Toutes ces classifications et typologies proposent d'étudier l'orthographe à différents niveaux et à partir de différents prismes en utilisant des terminologies différentes. Afin de décrire les graphèmes, nous utiliserons la terminologie utilisée par Catach. Un graphème impliqué dans la chaine sonore sera donc appelé un phonogramme. Un graphème portant des indices morphologiques sera appelé un morphogramme. Un graphème permettant de distinguer des homophones et incluant un potentiel « idéographique » est appelé logogramme ou « figure de mot ».

Cependant, nous n'emploierons pas la terminologie de Catach pour désigner les erreurs. En effet, nous n'utiliserons pas la distinction entre erreurs phonétiques et erreurs phonogrammique. Cette typologie différencie des erreurs telles **moner* / *mener*, dites phonétiques, des erreurs de diacritiques, placées sur le même plan que les erreurs n'altérant pas la valeur phonique. Cette distinction part de l'hypothèse que les erreurs phonétiques sont la conséquence possible d'une mauvaise perception du son et qu'il s'agit donc d'erreurs extragraphiques. Pour notre part, nous pensons qu'il peut effectivement s'agir d'erreurs dues à une mauvaise perception des sons, mais qu'il peut également s'agir d'une variation de prononciation (accent régional ou étranger), d'un défaut de catégorisation des sons en phonèmes ou encore d'une mauvaise connaissance des graphèmes et de leur emploi ou des correspondances phonographiques. En l'absence du scripteur, il ne nous appartient pas de choisir parmi ces différents cas de figures. Nous plaçons donc toutes les erreurs altérant la valeur phonique, y compris les erreurs de signes diacritiques, dans une seule catégorie. Nous appelons ces erreurs : erreurs de code phonographique, selon la typologie proposée par P.

Cappeau et M.-N. Roubaud (2005). En effet, nous considérons que dans la forme « donbre » (**1953**, *tombe*), le principe phonographique n'est pas respecté, que la cause se situe au niveau phonique (surdité, accent) ou au niveau phonographique (mauvaise connaissance des correspondances graphème-phonèmes).

Il faut entendre ici erreurs phonographiques au sens d'erreurs dans la correspondance phonie-graphie et qui impliquent des altérations dans la chaine phonique. Ainsi, écrire « tonb » (**1358**, *tombe*) n'entrave aucune correspondance graphème-phonème, nous considérons donc que cette forme est correctement graphiée d'un point de vue phonographique. En revanche, elle contrevient aux principes orthographiques. Ces derniers stipulent, en effet, que l'emploi d'une lettre *b* contraint la substitution de la lettre *n,* qui la précède, à la lettre *m*. Ils imposent également la présence de la lettre *e* après une consonne finale prononcée. Il s'agit donc d'une erreur de sélection des correspondances phonographiques selon le principe orthographique. Nous nommons donc ces erreurs : erreurs de sélection orthographique, à l'instar de P. Cappeau et M.-N. Roubaud (2005).

Afin d'élaborer un outil d'annotation capable de détecter et d'annoter des erreurs analysables en linguistique, il nous faut, au préalable, élaborer un schéma d'annotation répondant à ce besoin. Il faut donc que notre schéma puisse être automatisé, mais surtout il faut qu'il se fonde sur une certaine réalité linguistique, et donc choisir un niveau et un prisme d'analyse, afin de répondre au besoin de la description linguistique.

CHAPITRE 6 - SCHÉMA D'ANNOTATION

L'objet de ce mémoire est l'étude des apports du TAL à la constitution d'un corpus scolaire en vue d'exploitations didactiques et linguistiques. Cela signifie, entre autres, de marquer un premier pas vers une annotation automatique d'erreurs produites par des élèves de l'école primaire. Par conséquent, le modèle qui permettra cette annotation est l'élément central de cette étude. Il doit permettre de préciser et catégoriser les erreurs susceptibles d'être rencontrées dans le corpus. C'est sur ce modèle que se fondera par la suite l'exploitation du corpus. Nous allons alors réaliser un schéma décrivant les différentes erreurs, ce schéma doit remplir deux conditions. Il doit être :

– **descriptif** : il doit permettre une description linguistique des productions d'élèves répondant aux besoins des chercheurs ;

– **opératoire** : il doit permettre une annotation automatique des erreurs en fonction des possibilités du TAL.

Mais réaliser un tel schéma soulève de nombreuses questions, à commencer par : quelles erreurs annotons-nous ? Quelle méthodologie employons-nous ?

6.1. DÉLIMITATION DE LA NOTION D'ERREUR

Tout d'abord, il est nécessaire de définir cette notion d'***erreur***. En raison de l'emploi de l'informatique, nous avons choisi de définir le mot erreur au sens d'écart à la ***norme***, comme il est souvent d'usage dans l'apprentissage de l'orthographe (David, 2006). En effet, il est souvent plus facile d'analyser une production en termes d'écart à des formes normées et donc connues par les outils de TAL. Cependant, J. David (2006) nous met en garde contre cette conception puisqu'elle entraine une analyse des productions d'un élève en termes négatifs (erreurs, manques, écarts) et non en termes de connaissances déjà acquises. C'est pourquoi, dans la suite de notre travail, chaque unité d'étude se verra attribuée un statut (normé, erreur phonographique, etc.) qui peut être vu à la fois comme degré d'erreur, mais également comme un niveau d'acquisition.

6.1.1. DÉFINITION DE LA NORME

Parmi les différentes définitions et approches possibles de la notion de norme, nous avons distingué deux approches pour notre corpus : la normalisation est soit considérée comme une correction, dans ce cas notre approche à la norme se limite à ce qui est attendu

d'un élève de fin de CP, ou alors la normalisation considère tous les phénomènes susceptibles de faire l'objet de recherches ultérieures. Dans le premier cas, il serait envisageable de demander à plusieurs enseignants de CP de corriger quelques productions afin d'avoir une meilleure vision des phénomènes corrigés et non corrigés. Dans le second cas, une discussion avec les linguistes responsables du projet est nécessaire afin de déterminer au mieux les phénomènes concernés.

Nous avons choisi d'adopter la deuxième approche qui présuppose de faire des choix parmi les phénomènes en présence pour déterminer ceux que nous choisissons d'inclure dans la norme. Une production peut être normalisée à différents niveaux : stylistique, sémantique, syntaxique, lexical. Nous ne normaliserons pas chacun de ces niveaux, notamment parce qu'il s'agit parfois plus de choix d'interprétation que de véritables erreurs. Nous nous sommes par exemple demandé si la répétition du connecteur *et* pouvait être considérée comme une erreur ou tout au moins une maladresse. Mais nous avons fait le choix de ne pas en tenir compte.

> « le chat quoure é il tonbe // é il plére é sa maman / lui faiun qualin é le chat / na plumale » (**573**, *Le chat court et il tombe et il pleure et sa maman lui fait un câlin et le chat n'a plus mal.*)

Nous ne normalisons la segmentation en phrases que dans les cas d'absences de segmentation.

> « le chat march il sai fai mal / il pler sa maman vin le / chairchai » (**1577**, *Le chat marche. Il s'est fait mal. Il pleure. Sa maman vint le chercher.*)

Étant donné que nous nous intéressons à l'orthographe, tous les écarts à la norme orthographique, c'est-à-dire les écarts de graphie des mots, sont considérés comme des erreurs. De même, les blancs visibles en production et ne correspondant pas aux frontières de mots en langue française, comme « a prè » (**680**, *après*), ainsi que les frontières absentes, à l'exemple de « latrape » (**1116**, *l'attrape*), seront également considérés comme tels et seront désormais appelées *erreurs de segmentation*. Pour ces deux types d'erreurs, la norme adoptée correspond à la norme attendue pour des scripteurs adultes.

Le choix a également été fait de considérer certains phénomènes syntaxiques comme des erreurs, à l'exemple de l'absence de la négation ou du pronom *qui*.

> « un chat arété pas de se faire / mal. [...] » (**2977**, *Un chat n'arrêtait pas de se faire mal.*)

Nous ne développerons pas ici tous ces choix, un exposé détaillé est disponible en annexe (annexe 8). Nous retiendrons seulement que le système final devra prendre en compte différents niveaux d'erreurs.

6.1.2. SÉLECTION DES ERREURS

Parmi toutes les erreurs rencontrées dans le corpus, certaines nous paraissent plus pertinentes à traiter que d'autres. En effet, comment évaluer les erreurs syntaxiques d'une production sans avoir au préalable identifié les mots qui la composent ? D'autant qu'au CP les phrases ne sont pas encore clairement identifiées et leur syntaxe est souvent très approximative. Afin de résoudre ce problème d'identification des mots, il nous parait essentiel d'annoter en premier les erreurs qui se concentrent principalement sur les mots, à savoir les erreurs d'orthographe et les erreurs de segmentation. Pour que les phénomènes non annotés puissent tout de même faire l'objet d'une recherche par les utilisateurs du corpus, une proposition de réécriture « normée », selon la norme définie précédemment, sera proposée (cf. annexe 8).

6.2. ÉLABORATION DU SCHÉMA D'ANNOTATION

Une fois les types d'erreurs à étudier définis, il nous a fallu élaborer une méthodologie pour la réalisation du schéma d'annotation. L'approche adoptée se voulait empirique et itérative. Nous sommes ainsi partie du corpus, de son observation et de nos connaissances « naïves »[4] sur l'orthographe pour fonder notre analyse. Pour ce faire, nous nous sommes appuyé sur les premières observations réalisées à partir d'un sous-corpus de 17 productions et synthétisées sous formes de tableaux (cf. 2.4.). Ces tableaux présentant différents phénomènes répartis en niveaux d'analyse, ont constitué le point de départ de notre schéma d'annotation.

Après élaboration d'une première version, ce schéma a ensuite pu être testé sur de nouvelles productions (non incluses dans les 17 du sous-corpus). Lors de ces phases de tests, les productions ont été annotées manuellement en respectant le schéma d'annotation. Ces tests ont fait émerger de nouveaux phénomènes. Le schéma a alors été adapté à ces nouvelles productions, puis confronté aux modèles orthographiques et aux grilles d'erreurs développés précédemment (cf. chapitres 4 et 5). Cette comparaison a de nouveau donné lieu à certaines

[4] Connaissances qu'a toute personne ayant reçu un enseignement de l'orthographe française durant sa scolarité.

modifications. Enfin, ce nouveau schéma a pu être validé au moyen d'un nouveau test d'annotation de productions.

6.2.1. UNITÉ D'OBSERVATION

Comme nous l'avions mentionné lors de l'élaboration des premiers tableaux d'analyses, nous avions conservé les unités d'observations les plus englobantes possible afin de ne laisser de côté aucun phénomène. Les unités présentées alors nécessitent désormais d'être précisées et redéfinies afin de pouvoir servir à l'élaboration d'un schéma d'annotation. Il est nécessaire que chaque erreur soit portée par une unité d'analyse identifiée et unique.

6.2.1.1. LA SEGMENTATION

Pour étudier la segmentation, nous étudierons les segments visibles dans les productions et plus précisément le découpage que font les enfants de ces segments. Rappelons que le terme **segment** désigne toute séquence de lettres séparées par des frontières de mots, tandis que le terme **forme** désigne toute séquence de lettres séparées par des frontières de mots telles qu'on en trouve dans les lexiques de formes fléchies.

6.2.1.2. L'ORTHOGRAPHE

Nous avons commencé l'étude de l'orthographe en prenant pour unité la lettre. Cette première approche s'approchait du paradigme où le graphème est la lettre. Cependant, il est rapidement apparu que prendre la lettre comme unité de mesure ne permettait pas de traiter aisément la dimension phonographique de notre système d'écriture (cf. 2.3.). Un tel constat nous a amenée à transporter notre unité d'étude de la lettre au graphème. Pour utiliser la notion de graphème en TAL, il est nécessaire d'en donner une définition fonctionnelle, qui permette une identification des graphèmes de manière automatique.

Comme nous venons de l'évoquer, il nous faut élaborer une définition qui inclut la dimension phonographique de l'écriture. Nous définissons donc le graphème comme la lettre ou la séquence de lettres correspondant à un phonème ou à une absence de phonème dans la représentation phonologique de la forme attendue. Ainsi, toutes les lettres adjacentes qui permettent de transcrire un seul phonème correspondent à un graphème, de même que toutes les lettres adjacentes qui n'ont aucun correspondant phonique. Ainsi, par exemple, dans la forme *chat* nous distinguons 3 graphèmes : *ch*, *a* et *t*. Il n'est pas tenu compte de la morphologie pour le moment.

6.2.2. PRÉSENTATION DU SCHÉMA PAR NIVEAUX DE TRAITEMENTS

Le schéma que nous élaborons se traduira au niveau informatique par un format XML[5] dans lequel les erreurs sont identifiées et décrites par des attributs sur ces balises. Il faut donc que ce soit un schéma hiérarchique. Chaque niveau hiérarchique correspondra à des niveaux de traitements permettant d'annoter différents types d'erreurs. La présentation suivante exposera à la fois le raisonnement suivi, les différents niveaux contenus dans le schéma et les balises permettant de les annoter.

6.2.2.1. SÉQUENCE DE SEGMENTS

Les erreurs de segmentation sont portées par les segments. Cependant ce type d'erreur porte toujours sur plusieurs segments à la fois. Reprenons les exemples :

- « à près » (**1156**, *après*), ici l'erreur est portée conjointement par le segment « à » et « près » ;
- « latrape» (**1116**, *l'attrape*), cette erreur n'est portée que par un segment, mais correspond à trois formes différentes au niveau de la norme.

Il nous faut donc envisager de traiter ces erreurs au niveau des séquences de segments et non des segments isolés. Pour ce faire une balise <segmentation> est créée. Cette balise contient également un attribut permettant de distinguer hypersegmentation et hyposegmentation.

L'hyposegmentation est l'erreur qui consiste à agglutiner une ou plusieurs formes. La balise <segmentation type="hyposegmentation"></segmentation> permet d'annoter ces erreurs : « latrape » est annoté « <segmentation type="hyposegmentation">latrape</segmentation> ».

À l'inverse, l'hypersegmentation consiste à segmenter une forme en deux ou plusieurs segments. Pour l'annoter, la balise <segmentation type="hypersegmentation"/> sera notée à l'endroit de chaque coupure fautive, afin d'identifier l'endroit exact de la segmentation insérée. Par exemple, « à près » est annoté « à<segmentation type="hypersegmentation"/>près ».

6.2.2.2. SEGMENTS

Il nous a ensuite paru important d'annoter les erreurs au niveau du segment. Nous avons attribué trois statuts différents aux segments, selon leur écart à la norme et

[5] eXtensible Markup Language (http://www.w3.org/TR/xml/)

particulièrement selon leur rapport aux formes graphiques et phoniques de la forme normée attendue :

– **normé**, dans le cas où il est identique à la forme attendue, par exemple le segment « chat » (**1666**). Ces segments seront annotés au moyen de la balise : « chat ».

– **erroné (substitué) à phonologie normée**, lorsque la représentation phonologique qui peut être faite à partir des règles usuelles de lecture du français est identique à la représentation phonologique de la forme attendue, comme l'exemple « tonb » (**1596**, *tombe*). À l'instar de P. Cappeau et M.-N. Roubaud (2005) (cf. 5.3.4.), nous appellerons ces erreurs, *erreurs de sélection orthographique*, puisque nous considérons qu'il s'agit d'une mauvaise sélection de la forme graphique par rapport au contexte et à la norme. Cette erreur sera annotée : « tonb ».

– **erroné (substitué) à phonologie non normée** lorsque ni la graphie ni la représentation phonologique qui en découle ne correspondent à la forme attendue, à l'exemple de « dar » (**3006**, *part*). Nous considérons que ces erreurs sont le fruit d'une mauvaise sélection du matériau graphique en lien avec la forme sonore, elles reflètent donc une maitrise partielle du lien phonie-graphie, et seront appelées *erreurs de code phonographique*. Elles seront annotées comme dans l'exemple : « dar ».

6.2.2.3. GRAPHÈMES

Si les erreurs sont reportées au niveau du segment, la majorité des erreurs cependant prennent leur source au niveau du graphème. Ce niveau est donc fondamental dans notre schéma.

Toutefois, lorsque le segment est normé, il n'est nul besoin de descendre à un niveau inférieur. Dans les autres cas, tous les graphèmes du segment non normé seront décrits. Outre les trois statuts décrits plus hauts, à savoir : normé (« a » dans « chat », **1666**), substitué orthographiquement (« on » dans « tonb », **1596**, *tombe*) et substitué phonographiquement (« d » dans « dar », **3006**, *part*), un graphème peut également être omis (« e » dans « tonb », **1596**, *tombe*) ou inséré (« e » dans « soire », **3066**, *soir*).

Un segment peut également être omis ou inséré, cependant ces phénomènes ne seront pas traités ici, car ils relèvent de la syntaxe et non plus de l'orthographe ou de la segmentation.

« il ettai tune fou un peti chat et tai sor. ti […] » (**1228**, *Il était une fois un petit chat* ***qui*** *était sorti*).

Toutefois, ces cinq états ne suffisent pas à décrire l'ensemble des phénomènes d'erreur portés par les graphèmes. Mais surtout, ils ne permettent pas d'analyser le graphème selon notre optique de départ, c'est-à-dire en termes d'altération de la valeur phonique. Pour l'heure, la balise permettant d'annoter les graphèmes ressemble à la balise utilisée pour les segments : <grapheme ecart="normé"></grapheme>. Cependant, pour réaliser une analyse plus fine des erreurs, il est nécessaire d'y ajouter d'autres attributs.

Sonorité et valeur sonore

Si certains graphèmes permettent de transcrire la chaine sonore, d'autres n'ont aucun lien avec celle-ci. Il nous parait important de distinguer ces deux types de graphèmes. Nous les avons donc classés en deux grandes catégories : **sonores** ou **muets**. Nous pouvons éventuellement décliner la catégorie sonore en trois sous-catégories : **voyelle**, **consonne** et **semi-voyelle**. Il nous semble en effet que, même si nous ne l'avons pas pris en compte dans notre propre étude, cette différenciation peut être intéressante pour des études ultérieures. Cela permettra notamment de distinguer les phonèmes les plus susceptibles d'erreurs lors de leur transcription. Prenons l'exemple de la forme *part* : *p* et *r* sont des consonnes, *a* est une voyelle, *t* est un graphème muet.

Cette information sera portée par un nouvel attribut nommé sonorité. Un deuxième attribut sera ajouté, permettant de préciser la valeur sonore normée des phonogrammes. Cette valeur est écrite au format LIA (Béchet, 2001), format qui permet d'écrire un phonème à l'aide d'une suite de deux caractères non soumis aux problèmes d'encodage (cf. annexe 9 pour le détail de ce format).

Un segment comme « dar » (**3006**, *part*) sera annoté :

```
<grapheme ecart="phonographique" sonorite="consonne" phoneme="pp">d</grapheme>
<grapheme ecart="normé" sonorite="voyelle" phoneme="aa">a</grapheme>
<grapheme ecart="normé" sonorite="consonne" phoneme="rr">r</grapheme>
<grapheme ecart=" omis " sonorite="muet"></grapheme>
```

Fonction

Comme nous l'avons évoqué, les phénomènes grammaticaux inaudibles comme les marques de flexion ou d'accord semblent souvent absents dans ce corpus. Il nous semble donc intéressant de distinguer les graphèmes qui sont du domaine lexical des graphèmes qui

relèvent de phénomènes grammaticaux, particulièrement pour les graphèmes muets. Les graphèmes seront donc également divisés en deux fonctions selon que ce sont des **graphèmes lexicaux** ou des **graphèmes grammaticaux**. Est appelé graphème lexical tout graphème présent dans le radical, ainsi dans la forme *chats* : *ch*, *a* et *t* sont des graphèmes lexicaux, *s* est un graphème grammatical, tout comme dans *réveillent* : *r*, *é*, *v, e* et *ill* sont des graphèmes lexicaux, tandis que *ent* est un graphème grammatical.

Selon la définition des graphèmes faite précédemment, nous considérions *ts* dans la forme *chats* comme un graphème unique (« toutes les lettres adjacentes qui n'ont aucun correspondant phonique »). Ce qui ne manque pas de poser problème, il nous faut donc modifier notre définition. Il faut désormais ajouter une condition à cette définition : toutes les lettres adjacentes qui n'ont aucun correspondant phonique et qui ont même fonction. La suite *ts* est désormais considérée comme deux graphèmes différents *t* et *s*. L'exemple précédent (« dar », **3006**, *part*) sera désormais annoté :

<grapheme ecart="phonographique" fonction="lexicale" sonorite="consonne" phoneme="pp">
d</grapheme>
<grapheme ecart="normé" fonction="lexicale" sonorite="voyelle" phoneme="aa">a</grapheme>
<grapheme ecart="normé" fonction="lexicale" sonorite="consonne" phoneme="rr">r</grapheme>
<grapheme ecart=" omis " fonction="grammaticale" sonorite="muet"></grapheme>

Écart à la norme et valeur normée

Il nous semble également nécessaire, particulièrement dans les cas de substitution et d'omission, d'ajouter un attribut permettant de donner la valeur du graphème normé. Le *t* omis dans « dar » (**3006**, *part*) pourra désormais être annoté :

<grapheme ecart=" omis " fonction="grammaticale" sonorite="muet" valeur="t" ></grapheme>

Au niveau du graphème, on retrouve l'opposition entre phonologie normée et non normée selon que le graphème transcrit le phonème attendu ou non. Cependant, il nous semble intéressant d'affiner quelque peu cette opposition. Les graphèmes que nous reconnaissons comme substitués à phonologie normée (substitution orthographique) ne seront pas reconnus ainsi par tous les systèmes. En effet, rappelons que nous avons fait le choix de considérer un système phonologique minimal où /o/ et /ɔ/ sont regroupés sous un unique phonème, de même que /ɛ̃/ et /œ̃/, etc. Le segment « kopun » (**1138**, *copains*) a donc pour représentation phonologique /kopœ̃/, de même que sa forme normée, alors que dans un système plus expansif sa représentation phonologique serait /kopœ̃/, tandis que celle de sa forme normée serait /kopɛ̃/. Ces cas seront donc distingués des cas comme « tonb » (**1596**, *tombe*) qui présente une différence purement graphique avec sa forme normée, quel que soit le

système phonologique considéré. Nous nommerons ces nouveaux cas : **substitutions orthographiques à phonème étendu**, abrégées en **substitutions orthographiques étendues**. Précisons que cette erreur ne peut porter que sur des voyelles puisque nous n'avons supprimé d'oppositions qu'entre voyelles. Précisons également que les phonèmes concernés seront transcrits par la valeur de l'archiphonème. Les graphèmes *é* et *ai* auront tout deux pour valeur sonore /ei/.

Le segment « kopun » sera donc annoté :

<grapheme ecart="orthographique" fonction="lexicale" sonorite="consonne" phoneme="kk" valeur="c">k</grapheme>
<grapheme ecart="normé" fonction="lexicale" sonorite="voyelle" phoneme="oo">o</grapheme>
<grapheme ecart="normé" fonction="lexicale" sonorite="consonne" phoneme="pp">p</grapheme>
<grapheme ecart="orthographique étendue" fonction="lexicale" sonorite="voyelle" phoneme="in" valeur="ain">un</grapheme>
<grapheme ecart="omis" fonction="grammaticale" sonorite="muet" valeur="s"></grapheme>

Parmi les graphèmes présents en corpus et ne transcrivant pas le phonème attendu, il nous semble qu'il y ait là aussi des distinctions à faire. Et pour cause, certaines de ces erreurs ne sont pas dues à une méconnaissance du code phonographique mais plutôt à une méconnaissance des règles d'association des graphèmes. Nous illustrerons notre propos à l'aide de l'exemple « ce » (**1064**, *que*). Dans cet exemple, écrire le graphème *c* pour transcrire le phonème /k/ ne relève pas d'une erreur de code phonographique puisque ce graphème est un des correspondants phonographiques de /k/, mais dans ce contexte, il ne l'est pas. C'est donc le contexte qui sélectionne la valeur du graphème et l'on peut appeler ce type d'erreurs des erreurs de **valeur du graphème**. Le segment « ce » sera annoté :

<grapheme ecart="valeur" fonction="lexicale" sonorite="consonne" phoneme="kk" valeur="qu">c</grapheme>
<grapheme ecart="normé" fonction="lexicale" sonorite="voyelle">e</grapheme>

Prenons un autre exemple : « revil » (**1666**, *réveille* prononcée /rEvEj/), prononcé /rŒvil/. La notion de graphème étant définie par rapport à la forme attendue, les deux lettres *i - l* sont considérées comme un graphème unique, ce qui pose ici deux problèmes. En premier lieu, il remplace le graphème *ill*, il s'agit donc d'une substitution de graphème, mais il présente également la particularité de ne pas se prononcer comme la forme normée, sans qu'il s'agisse d'un problème de graphie du graphème. En effet, remplacé par sa forme normée *ill*, il serait également prononcé /il/ et non /j/. Le problème vient ici non pas du graphème lui-même mais de l'absence d'un graphème voisin qui dispose d'une valeur auxiliaire (cf. Blanche-Benveniste et Chervel, 1969), en l'occurrence *e*. Se pose alors la question de savoir à quel

niveau l'erreur doit être annotée : doit-elle être annotée sur les deux graphèmes ou sur un des deux graphèmes uniquement ? Le but de notre annotation étant d'annoter la graphie et non la capacité à lire selon les règles de combinaison des graphèmes, nous choisissons de ne pas reporter l'erreur sur le graphème *ill*. L'erreur sera donc portée par le graphème à valeur auxiliaire. Pour ce faire, la mention **auxiliaire** pourra donc être ajoutée à la sonorité. Le segment « revil » sera ainsi annoté :

```
<grapheme ecart="normé" fonction="lexicale" sonorite="consonne" phoneme="rr">r</grapheme>

<grapheme ecart="phonographique" fonction="lexicale" sonorite="voyelle" phoneme="ei"
valeur="é">e</grapheme>
<grapheme ecart="normé" fonction="lexicale" sonorite="consonne" phoneme="vv">v</grapheme>
<grapheme ecart="omis" fonction="lexicale" sonorite="voyelle auxiliaire" phoneme="ei"
valeur="e"></grapheme>
<grapheme ecart="orthographique" fonction="lexicale" sonorite="semi-voyelle" phoneme="yy"
valeur="ill">il</grapheme>
<grapheme ecart="normé" fonction="grammaticale" sonorite="muet">e</grapheme>
```

Il nous faut également annoter un phénomène relativement fréquent, il s'agit de la transcription de la liaison sur le mot suivant comme dans « des sotre » (**1156**, *des autres*) ou encore « il zon desider » (**667**, *ils ont décidé*). Certains auteurs (Roubaud et Cappeau, 2005) l'analysent comme une erreur de segmentation. Cependant, la présence double de la consonne comme c'est le cas dans le premier exemple nous en empêche. Nous considèrerons ce phénomène comme une insertion de graphème. Toutefois, comme il n'a ni une fonction lexicale, ni une fonction grammaticale mais une fonction phrastique, nous ajouterons une troisième fonction nommée **liaison** pour annoter cette erreur spécifique. Cette erreur ne portera donc que sur l'insertion de consonne. Le segment « zon » s'annotera ainsi :

```
<grapheme ecart="inséré" fonction="liaison" sonorite="consonne" phoneme="zz">z</grapheme>
<grapheme ecart="normé" fonction="grammaticale" sonorite="voyelle" phoneme="on">on
</grapheme>
<grapheme ecart="omis" fonction="grammaticale" sonorite="muet" valeur="t"></grapheme>
```

Les accents ne sont pas traités comme des signes diacritiques mais les lettres accentuées sont considérées comme des graphèmes à part entière. Ce choix s'explique par la grande fréquence d'opposition telle que *ai* remplacé par *é* ou *è*, à raison de 10 occurrences sur un corpus de 17 productions. Or, pour expliquer une telle substitution en considérant l'accent comme signe diacritique, il faudrait considérer la substitution de *e* à *ai*, substitution qui modifie la phonologie puis l'ajout d'un signe diacritique, on considère donc deux erreurs et non une. Dans l'exemple « il eté diféran » (**1156**, *Il était différent*), le segment « eté » sera annoté :

<grapheme ecart="phonographique" fonction="lexicale" sonorite="voyelle" phoneme="ei" valeur="é">e</grapheme>
<grapheme ecart="normé" fonction="lexicale" sonorite="consonne" phoneme="tt">t</grapheme>
<grapheme ecart="orthographique étendue" fonction="grammaticale" sonorite="voyelle" phoneme="ei" valeur="ai">é</grapheme>
<grapheme ecart="omis" fonction="grammaticale" sonorite="muet" valeur="t"></grapheme>

Il nous semble important également de préciser que les consonnes doubles sont considérées comme un seul graphème, l'omission ou l'insertion d'une des consonnes est considérée comme une substitution qui peut avoir différents statuts selon qu'elle modifie la valeur phonique (« aprisse», **1986**, *a pris)* du graphème ou non (« diféran », **1156**, *différent*).

6.2.2.4. LETTRES

La majorité des phénomènes sont envisagés au niveau du graphème, ainsi l'écriture du graphème *on* lorsque le graphème *om* est attendu est considéré comme une substitution de graphème. Toutefois, il nous parait plus difficile de considérer le graphème *n* dans « mamn » (**1346**, *maman*) comme une substitution du graphème *an*. Il nous faut considérer un niveau inférieur c'est-à-dire la lettre. L'erreur présentée ci-dessus est alors considérée comme une omission de lettre dans un graphème digraphe ou trigraphe. Précisons que nous ne partageons pas le point de vue des auteurs C. Blanche-Benveniste et A. Chervel (1969) quant à la valeur auxiliaire ou la valeur zéro de *a* dans la séquence *ain*, ni de *r* et *t* dans les séquences *er* et *et*. L'absence de la lettre *e* sera donc considérée comme une omission de lettre et non de graphème.

En revanche, de même que nous avons considéré la substitution de *il* à *ill* (cf. 6.2.3.) comme une substitution de graphèmes, nous considèrerons la substitution de *in* à *ain*, et inversement, comme une substitution de graphèmes. Notre critère pour distinguer ces cas des cas précédents est qu'ils sont substitués à un graphème existant et à même valeur phonique.

La lettre n'étant pas notre unité principale d'étude, seuls la valeur de la norme et l'écart à celle-ci seront précisés, selon les cinq statuts principaux : normé, omis, inséré, substitué orthographiquement (valeur phonique conservée), substitué phonographiquement (valeur phonique altérée, cf. 6.2.2.2. et 6.2.2.3.).

Le graphème « n » (*an*) de « mamn » (**1346**, *maman*) sera donc annoté :

```
<grapheme ecart="phonographique" fonction="lexicale" sonorite="voyelle" phoneme="an"
valeur="an">
        <lettre ecart="omis" valeur="a"></lettre>
        <lettre ecart="normé">n</lettre>
</grapheme>
```

6.2.2.5. SÉQUENCES DE GRAPHÈMES

Enfin, certains phénomènes ne peuvent s'appréhender qu'à un niveau supérieur au graphème, sans englober la totalité du segment. Il s'agit, par exemple, des **inversions** de graphèmes, à l'exemple de « apér » (**3006**, *après*).

Prenons pour exemple la production « avec c'est chaton » (**1301**, *avec ses chatons*). Ici, la forme *ses* est remplacée par la suite de segments « c'est ». Il ne s'agit pas d'un segment normé, donc selon la logique donnée plus haut nous allons en étudier chacun des graphèmes. Cependant, on comprend bien que cela n'aurait pas de sens, il s'agit d'un mot remplacé par un autre mot ou séquence de mots connue par l'enfant. La sélection lors de la mise à l'écrit ne s'est donc pas faite graphème par graphème mais à un niveau plus global. Nous ne placerons pas cette erreur au niveau du segment parce que l'on peut trouver ce processus en milieu de mot, comme dans les exemples « etnorme » et « estnorme » (**1289**, *énorme*). De plus, nous ne considèrerons pas cette erreur comme une erreur de segmentation puisque c'est l'ensemble qui est considéré par l'enfant et non chaque segment séparé. Pour désigner ce phénomène, nous reprendrons le mot employé par N. Catach (1978), il s'agit de **figures de mots** ou **logogrammes**. Le segment « c'est » (*ses*) sera donc annoté :

```
<segGraph ecart="logographique" valeur="ses">c'est</segGraph>
```

Enfin, on retrouve des phénomènes non spécifiques à l'écrit et rares dans notre corpus, notamment la surgénération de règles, comme dans l'exemple « [...] ai se fesa très males [...] » (**1168**, *et se fit très mal*), où la surgénération des règles de conjugaison au passé simple produit une forme erronée. L'erreur peut alors se répercuter sur plusieurs graphèmes, nous appellerons ce type d'erreur des **substitutions flexionnelles**, annotées comme dans l'exemple « fesa » (*fit*) :

```
<grapheme ecart="normé" fonction="lexicale" sonorite="consonne" phoneme="ff">f</grapheme>
<segGraph eacrt="flexionnelle" valeur="it">esa</segGraph>
```

Nous parlons de séquence de graphèmes mais nous ne parlerons pas de syllabe parce que très peu de phénomènes se rapporte spécifiquement à cette structure particulière, le terme

de séquence de graphèmes est plus englobant. La syllabation pourra se faire par la suite, sans rentrer dans l'annotation.

6.2.2.6. SIGNES GRAPHIQUES

La catégorie des graphèmes n'incluant, à notre sens, que les caractères lettres, tous les autres signes ne pourront pas être traités dans cette catégorie, une catégorie signes graphiques est alors ajoutée au schéma d'annotation, incluant la ponctuation, les guillemets, les tirets, les apostrophes. De manière générale, nous ne traitons pas non plus la ponctuation. Néanmoins, dans certains cas comme « sor.ti » (**1228**, *sorti*) il parait important de signifier que le point est de trop. Les signes graphiques ne présentent pour attribut que l'écart à la norme qui peut prendre quatre valeurs : normé, omis, inséré et substitué.

Tous ces traitements ont permis de donner lieu au schéma suivant :

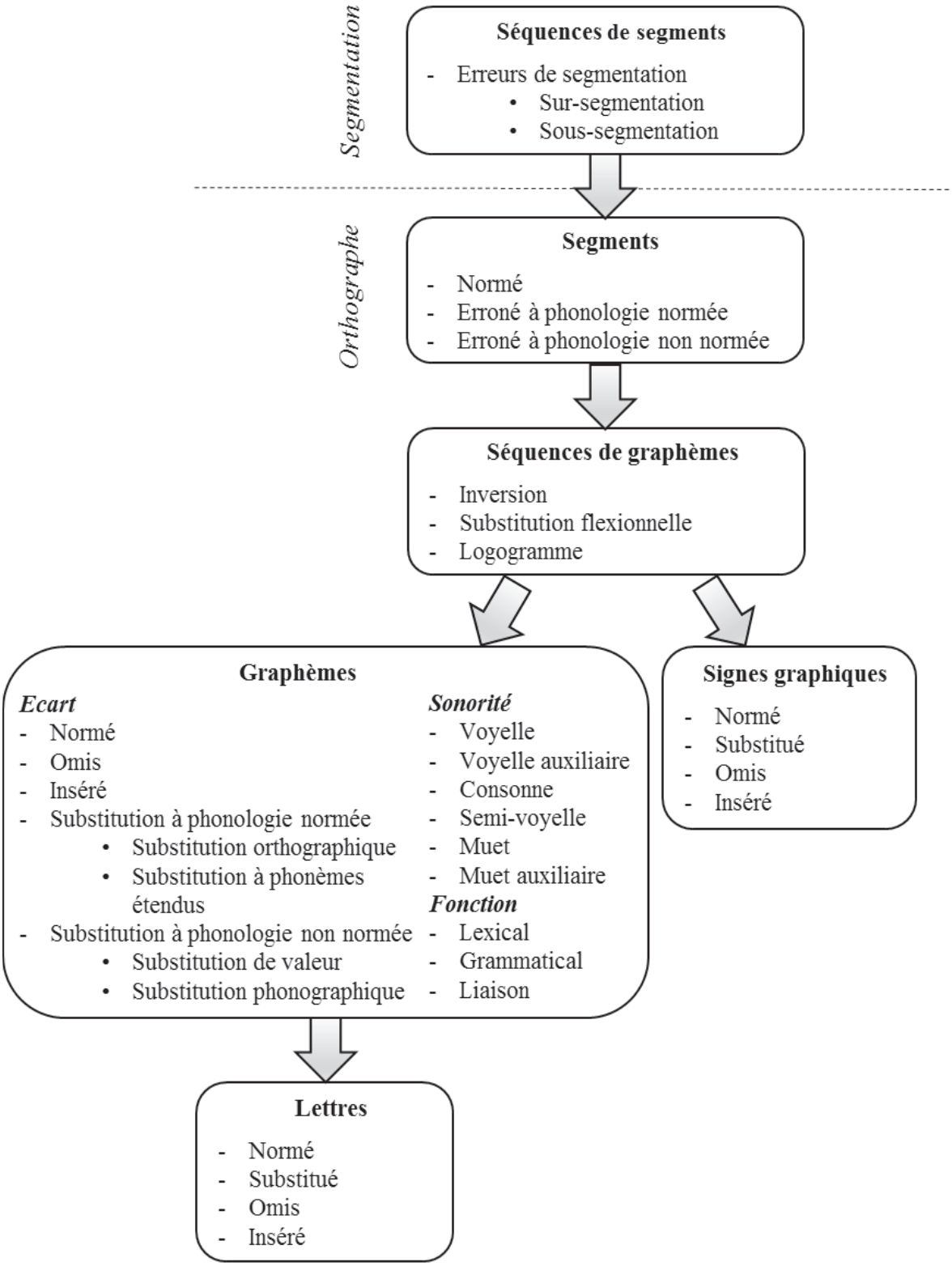


Figure 6. Schéma d'annotation des erreurs d'orthographe et de segmentation

Toutes les possibilités que permet notre schéma ne sont pas trouvées en corpus mais un certain nombre d'entre elles peuvent être exemplifiées à l'aide de celui-ci :

Niveau de l'erreur	Type d'écart			Exemple	Forme normée	N° de la production
Séquence de segments	*Hyposegmentation*			Listoir	L'histoire	3129
	Hypersegmentation			ton bèt	tombé	1301
Segments	*Normé*			chat	chat	1301
	Erroné à phonologie normée			fé	fait	3030
	Erroné à phonologie non normée			i	il	3031
Séquence de graphèmes	*Inversion*			ér (apér)	rè (après)	3006
	Substitution flexionnelle			esa (fesa)	it (fit)	1168
	Logogramme			c'est	ses	1301
Graphèmes	**Type d'écart**	**Fonction**	**Sonorité**			
	Normé	*Lexical*	*Voyelle*	a (cha)	a (chat)	3177
			Voyelle auxiliaire	e (reveilla)	e (réveilla)	1134
			Consonne	v (reveilla)	v (réveilla)	1134
			Semi-voyelle	ill (reveilla)	ill (réveilla)	1134
			Muet	t (chat)	t (chats)	1290
		Grammatical	*Voyelle*	a (reveilla)	a (réveilla)	1134
			Muet	e (tonbe)	e (tombe)	1350
	Omis	*Lexical*	*Voyelle*	(ch)	a (chat)	3006
			Voyelle auxiliaire	(revil)	e (réveille)	1666
			Consonne	(cheirei)	ch (chercher)	1346
			Muet	(ch)	t (chat)	3006
			Muet auxiliaire	(revil)	e (réveille)	1986
		Grammatical	*Voyelle*	(tomb)	é (tombé)	3026
			Muet	(tonb)	e (tombe)	1358
	Inséré	*Lexical*	*Voyelle*	e (pelere)	(pleure)	2976
		Grammatical	*Muet*	e (mongée)	(manger)	2986
			Muet auxiliaire	e (aprisse)	(a pris)	1986
		Liaison	*Consonne*	t (tune)	(une)	1602
	Substitution orthographique	*Lexical*	*Voyelle*	on (tonb)	om (tombe)	1358
			Voyelle auxiliaire	è (révèille)	e (réveille)	2911
			Consonne	f (diféran)	ff (différent)	1156
			Semi-voyelle	i (réveia)	ill (réveilla)	1956
			Muet	e (foie)	s (fois)	2022
		Grammatical	*Voyelle*	è (étè)	ai (était)	1956
			Muet	e (dore)	t (dort)	1346
	Substitution à phonèmes étendus	*Lexical*	*Voyelle*	un (kopun)	ain (copains)	1166
			Voyelle auxiliaire	é (révéi)	e (réveille)	2386
		Grammatical	*Voyelle*	é (eté)	ai (était)	1156
	Substitution de valeur	*Lexical*	*Voyelle*	e (reveier)	é (réveillé)	2883
			Consonne	c (ce)	qu (que)	1064
		Grammatical	*Voyelle*	e (fe)	ai (fait)	3006
	Substitution phonographique	*Lexical*	*Voyelle*	on (mongée)	an (manger)	2986
			Consonne	d (dar)	p (part)	3006
			Muet	m (maim)	s (mais)	1956
Lettres	*Normé*			n (mamn)	n (maman)	1346
	Omis			a (mamn)	a (maman)	1346
Signes graphiques	*Omis*			(Listoir)	' (L'histoire)	3129
	Inséré			. (sor.ti)	(sorti)	1228

Tableau 18. Exemples d'erreurs classées selon leur type

Les principales balises utilisées sont données dans un tableau (tableau 19), tandis que des exemples d'annotation de productions entières sont visibles en annexe (cf. annexe 10).

Niveau de l'erreur	**Type d'écart**			**Balise ou exemple de balise**
Séquence de segments	Hyposegmentation			<segmentation type="hyposegmentation">
	Hypersegmentation			<segmentation type="hypersegmentation">
Segments	Normé			
	Erroné à phonologie normée			
	Erroné à phonologie non normée			
Séquence de graphèmes	Inversion			<segGraph eacrt="inversion" valeur="">
	Substitution flexionnelle			<segGraph eacrt="flexionnelle" valeur="">
	Logogramme			<segGraph eacrt="logographique" valeur="">
Graphèmes	**Type d'écart**	**Fonction (exemples)**	**Sonorité (exemples)**	
	Normé	Lexical	Voyelle	<grapheme ecart="normé" fonction="lexicale" sonorite="voyelle" phoneme="">
	Omis	Grammatical	Voyelle auxiliaire	<grapheme ecart="omis" fonction="grammaticale" sonorite="voyelle auxiliaire" phoneme="" valeur="">
	Inséré	Grammatical	Muet	<grapheme ecart="inséré" fonction="grammaticale" sonorite="muet">
		Liaison	Consonne	<grapheme ecart="inséré" fonction="liaison" sonorite="consonne" phoneme="">
	Substitution orthographique	Lexical	Semi-voyelle	<grapheme ecart="orthographique" fonction="lexicale" sonorite="semi-voyelle" phoneme="" valeur="">
	Substitution à phonèmes étendus	Lexical	Voyelle	<grapheme ecart="orthographique étendue" fonction="lexicale" sonorite="voyelle" phoneme="" valeur="">
	Substitution de valeur	Grammatical	Voyelle	<grapheme ecart="valeur" fonction="grammaticale" sonorite="voyelle" phoneme="" valeur="">
	Substitution phonographique	Lexical	Consonne	<grapheme ecart="phonographique" fonction="lexicale" sonorite="consonne" phoneme="" valeur="">
Lettres	Normé			<lettre ecart="normé">
	Omis			<lettre ecart="omis">
	Inséré			<lettre ecart="inséré">
	Substitution orthographique			<lettre ecart="orthographique">
	Substitution phonographique			<lettre ecart="phonographique">
Signes graphiques	Normé			<signeGraphique ecart="normé">
	Omis			<signeGraphique ecart="omis">
	Inséré			<signeGraphique ecart="inséré">
	Substitué			<signeGraphique ecart="substitué">

Tableau 19. Exemples de balises selon le type d'erreur

Enfin, outre cette annotation, un deuxième niveau d'annotation est prévu à l'aide d'un étiqueteur morphosyntaxique permettant d'attribuer une catégorie grammaticale à chaque segment. Cet étiquetage, nécessaire à l'exploitation (interrogation) du corpus, aidera également au traitement des erreurs grammaticales.

6.3. CONCLUSION

Suite à une première observation du corpus, nous avons décidé de traiter quasi-exclusivement le versant phonographique de l'orthographe et très peu le versant sémiographique, contrairement aux modèles existants. En effet, il nous a semblé qu'en ce début d'apprentissage, le jeune scripteur avait pour principale préoccupation de rendre compte de la dimension orale de son message, se situant dans un traitement encore essentiellement alphabétique et peu orthographique (au sens de Fijalkow, cf. 5.1.3). Nous avons donc laissé de côté une grande partie des processus morphographiques.

Cependant, comme l'ont montré Martinet *et al.* (2004), les apprenants sont déjà capables d'emmagasiner et d'utiliser des connaissances lexicales après quelques mois d'apprentissage. Ce qui explique que certains mots fréquents, comme le mot *chat*, soient le plus souvent correctement orthographiés (à hauteur de 85% pour *chat*, cf. 2.1.). Néanmoins, lorsque plusieurs de ces mots ou groupes de mots faisant l'objet de procédure orthographique sont homophones, à l'exemple de *c'est* et *ses*, cela donne lieu à des confusions, ce que N. Catach appelle les erreurs logogrammiques ou encore « figures de mots ». Or, ces erreurs sont relativement nombreuses dans notre corpus, ce pan de la sémiographie est donc envisagé dans notre schéma.

Comme nous l'avons vu précédemment (cf. chapitre 5), au fur et à mesure de l'apprentissage, l'utilisation des procédures orthographiques et sémiographiques croît. Il sera donc nécessaire, dans une perspective de corpus longitudinal, d'ajouter une dimension morphologique à notre schéma.

Il nous faut maintenant vérifier que le schéma réalisé et les choix effectués soient applicables autant à la réalité du corpus qu'aux contraintes de la programmation.

CHAPITRE 7 - MISE EN APPLICATION : LES ERREURS DE SÉLECTION ORTHOGRAPHIQUE

7.1. INTRODUCTION

Le schéma élaboré précédemment doit permettre une annotation des erreurs à partir de procédés automatiques. Pour vérifier qu'il remplit cette condition, nous essayons ici une première analyse automatique, celle des erreurs de choix orthographique. Pour rappel, il s'agit d'erreurs qui ne portent pas atteinte à la valeur phonique du graphème. Ce sont des segments ayant la même représentation phonologique que leur forme normée comme le segment « peti » (**652**, réécriture normée : *petit*). Cependant, cette catégorie d'erreur étant très large, nous nous limiterons aux segments ne comportant que des erreurs de type orthographique pures, écartant les erreurs de substitution logographique (remplacement d'un mot ou d'un fragment de mot par un autre à même valeur phonique). Les erreurs que nous traitons relèvent des étiquetages et <grapheme ecart="orthographique"> dans notre schéma d'annotation.

7.1.1. MÉTHODOLOGIE

Ne traitant que les segments dont la phonologie n'est pas atteinte par la graphie de l'élève, nous postulons qu'il est possible de retrouver leur forme normée à partir de leur représentation phonologique. Après l'obtention de celle-ci, il nous faut donc la comparer aux représentations phonologiques d'un ensemble de formes normées. Cette méthode devrait permettre d'extraire une ou plusieurs formes dont la représentation phonologique est identique à celle du segment étudié, parmi lesquelles devrait se trouver la forme normée attendue en corpus.

7.1.2. HYPOTHÈSE DE TRAVAIL

Pour effectuer ce travail, nous partons de l'hypothèse que le contexte de production du corpus constitue un élément susceptible de faciliter nos analyses. Le fait que les scripteurs soient des élèves de CP permet de faire l'hypothèse que le lexique général et les structures utilisées seront relativement limités. De plus, les productions ont été écrites après présentation de quatre images et une partie non négligeable de ces écrits se rapproche d'une description de

ces images (cf. 2.3). Le vocabulaire y est donc relativement restreint et correspond aux éléments de l'image. Cette hypothèse sera utilisée et discutée par la suite.

7.1.3. VOCABULAIRE

Pour plus de clarté, les formes attestées en corpus, c'est-à-dire sur lesquelles s'effectuera le traitement seront appelées **formes** ou **formes erronées**. Les formes trouvées dans les lexiques, correspondant à des formes du français, seront appelées **formes normées**, tandis que les formes normées attendues dans le contexte du corpus seront désignées par les termes **formes attendues**.

7.2. OUTILS

Une analyse manuelle ne pouvant être effectuée sur tout le corpus, nous réutiliserons donc le sous-corpus des productions se terminant par 6. Notre corpus s'étant par ailleurs développé, ce sous-corpus contient dès lors 20 productions, soit 471 segments.

7.2.1. DÉTECTER LES ERREURS AVEC TREETAGGER

Le premier travail, préliminaire à toute analyse, consiste à repérer les erreurs présentes dans notre corpus. Il est d'usage en TAL de considérer comme erreur, tout segment non contenu dans un lexique de formes fléchies du français. Appliquée à notre corpus, cette méthode laisserait bien trop d'erreurs, notamment des erreurs logographiques (remplacement d'une forme par une forme homophone), morphographiques et de segmentation (une forme étant parfois segmentée en formes attestées en français). Il nous faudrait donc, dans l'idéal, trouver une autre méthode de détection. Cependant, identifier de telles erreurs demanderait des analyses à un niveau aussi bien sémantique que syntaxique, ce que nous ne sommes pas en mesure de fournir pour le moment.

Nous nous contenterons donc dans cette étude de détecter les erreurs à l'aide de l'outil d'étiquetage morpho-syntaxique TreeTagger (Schmid, 1994) utilisé précédemment (cf. 2.4.). Il présente l'avantage d'être libre d'utilisation. Toutefois, il est important de préciser que l'utilisation de TreeTagger implique que la définition de mot utilisée dans ce travail est contrainte par celle qu'en fait l'outil. Au sens de TreeTagger, un segment est une séquence de caractères séparée par un blanc graphique, un signe de ponctuation, ou un autre caractère graphique comme l'apostrophe, le guillemet ou encore le tiret.

Cet outil permet d'associer automatiquement à chaque forme une catégorie syntaxique et un lemme. Pour étiqueter, TreeTagger se base sur une liste de formes fléchies. Lorsqu'une forme n'est pas contenue dans cette liste, elle est donc considérée comme n'appartenant pas aux formes fléchies du français et l'étiquette <*unknown*> lui est attribuée en place du lemme. Nous considérons ces formes comme des formes erronées et les autres comme des formes normées. Ainsi, nous pouvons utiliser cette étiquette comme critère pour distinguer les formes que nous traiterons de celles que nous ne traiterons pas car jugées sans erreurs au sens de TreeTagger.

Naturellement, cette méthode n'est pas exacte. En effet, il peut y avoir des formes reconnues comme formes du français mais comportant des erreurs, comme l'absence d'un *s* marquant un pluriel, mais nous faisons l'hypothèse que les formes sélectionnées à l'aide de cette méthode sont toutes des formes erronées. Nous allons vérifier cette hypothèse avant de continuer notre analyse.

Sur les 471 segments qui composent notre sous-corpus, 118 sont étiquetés <*unknown*>, tandis que 353 segments ont pu être étiquetés. Notre méthode permet donc de reconnaitre un taux d'erreur de 25,1% dans notre corpus. Parmi ces 118 segments, 4 sont jugés erronés à tort. En effet, dans la production **576** « r r r r bousm miaou » (réécriture normée : *RRRR BOUM MIAOU*), TreeTagger reconnait « r r r r » comme quatre segments différents, tous inconnus. La lettre R est figurée quatre fois sur la première image, de même que les onomatopées BOUM et MIAOU MIAOU, on peut donc supposer que l'enfant a appliqué une procédure logographique, c'est-à-dire une mémorisation d'une image de mot qu'il a reproduit sur sa feuille (Frith, 1985), ce n'est donc pas à proprement parler une erreur. Même si ce corpus est trop restreint pour pouvoir en faire des statistiques fiables, il peut tout de même nous donner un ordre de grandeur que nous pouvons espérer approcher. Dans le cas présent, 118 erreurs ont été repérées, parmi lesquelles 114 en sont effectivement. Nous pouvons donc attendre de notre méthode que le rappel de notre méthode avoisine les 95% (96,7% sur 20 productions).

Parmi les 353 segments identifiés par TreeTagger, 51 comportent des erreurs sans être étiquetés <*unknown*>, dont :

– 42 impliquent une mauvaise reconnaissance du lemme : par exemple, « dore » (**1346**, *dort*) que TreeTagger reconnait comme une forme du verbe DORER et non DORMIR,

– 9 sont des erreurs qui ne modifient pas l'étiquetage que fait TreeTagger, ce sont des erreurs morphogrammiques. Par exemple, « les chaton » (**1336**, *les chatons*) est analysé comme

un nom dont le lemme est CHATON malgré l'absence du *s* pluriel. Sont également comptabilisées dans ces 9 formes les erreurs qui modifient l'étiquetage morphologique mais non syntaxique et lemmatique. Par exemple, dans « elle er tomber » (**1226**, *elle est tombée*), *tomber* est analysé comme le verbe TOMBER à l'infinitif. Le fait que c'est un verbe dont le lemme est TOMBER est exact, cependant il ne devrait pas être à l'infinitif mais au participe passé.

Si l'on prend en compte ces 51 erreurs, le rappel, c'est-à-dire le nombre d'erreurs trouvées parmi les erreurs de notre corpus, de notre méthode est de 69,3%. Dans notre étude, nous ne nous intéressons qu'aux erreurs d'orthographes lexicales et non aux erreurs de flexions. Dans ce cas, nous ne considérons plus que 42 erreurs et nous pouvons estimer que la précision de notre méthode avoisinera 70 à 75% (73,2% sur 20 productions).

Les erreurs repérées par cette méthode se composent exclusivement de formes inconnues en français, bien qu'étant des erreurs de sélection orthographique, les erreurs logographiques, comme « c'est » (**1301**, *ses*), ne seront donc pas repérées ici et notre analyse portera exclusivement sur le niveau du graphème.

7.2.2. PHONÉTISER AVEC LIA_PHON

Une fois les segments à traiter sélectionnés, il est nécessaire de les convertir pour obtenir leurs représentations phonologiques et ainsi pouvoir les comparer aux formes normées. Pour ce faire, nous utilisons l'outil LIA_PHON développé par F. Bechet (2001) au Laboratoire Informatique d'Avignon. Cet outil présente l'avantage d'être gratuit (licence GNU) et de permettre de phonétiser à la fois des segments isolés mais également des textes. De plus, étant développé à partir de règles et non d'un lexique (cf. 3.3.2.), il permet de « phonétiser » des mots inconnus ou mal orthographiés.

Après tokenisation et étiquetage du texte, cet outil propose une phonétisation du texte segment par segment selon différents formats : le format SAMPA ainsi qu'un format spécifique appelé format LIA (cf. 6.2.2.3.). Une syllabation des segments est également proposée.

Cependant, il est important de noter que LIA_PHON est initialement destiné à des applications en synthèse vocale. Il a donc été élaboré en vue de phonétiser des phrases voire des textes, et phonétise par conséquent en tenant compte du contexte dans lequel est placé le mot à traiter. Ainsi, le même mot peut avoir une représentation phonique différente selon qu'il

est présenté en contexte ou hors contexte. Le système va notamment prendre en compte les liaisons, comme dans « Le chat c'est éloinié de sa maman. » (**1556**, *Le chat s'est éloigné de sa maman*) où « éloinié » est transcrit /telwanje/ (/tteillwwaannyyei/ format LIA), tandis que, présenté isolément, il est transcrit /elwanje/ (/eillwwaannyyei/ format LIA). Cependant, nous n'utilisons pas cet outil à des fins de prononciation mais pour connaitre quels sons ont réellement été encodés par les jeunes scripteurs, il est donc nécessaire d'effacer tous ces phénomènes de lissage en présentant au système chaque forme isolément.

Il nous faut également tenir compte du fait que la segmentation peut différer entre TreeTagger et LIA_PHON, ainsi *parce que*, considéré comme deux formes distinctes par TreeTagger, est considéré comme une forme unique par LIA_PHON. De plus, chacun de ces outils propose un étiquetage syntaxique, toutefois les étiquettes utilisées ne sont pas les mêmes et ne donnent donc pas les mêmes résultats. TreeTagger étant un outil spécifique d'étiquetage morphosyntaxique, nous nous appuierons sur les étiquettes qu'il propose.

Pour la suite de ce travail, le format utilisé sera le format LIA. Même s'il est moins répandu que les formats API et SAMPA, il présente l'avantage de n'utiliser que les 26 lettres de l'alphabet, ne contenant aucun signe diacritique ou caractère particulier. De plus, le nombre de caractères utilisé pour encoder un phonème est fixe, en effet chaque phonème est représenté par deux signes graphiques. Ce format nous parait donc plus simple d'utilisation d'un point de vue informatique. Notons que pour le futur outil d'exploitation du corpus, il sera toujours possible de transposer le codage LIA en API par exemple.

7.3. ÉTUDE

7.3.1. IDENTIFIER LES ERREURS À TRAITER

7.3.1.1. MÉTHODE

Avant de pouvoir élaborer une méthode d'annotation des erreurs, il nous faut d'abord repérer manuellement dans notre corpus les formes qui répondent à nos critères d'études afin de les analyser plus spécifiquement. Pour ce faire, il est nécessaire de comparer la forme phonologique de chaque forme non reconnue à la forme phonologique de sa forme normée. Afin de faciliter cette étape, nous nous appuyons sur un lexique de formes fléchies du français « phonétisé ». Nous pouvons alors, pour chaque forme reconnue comme erronée, rechercher sa forme normée dans le lexique et comparer les représentations phonologiques des deux formes.

Nous utilisons la base de données Lexique 3, élaborée par B. New et C. Pallier en 2005, qui fournit près de 130 000 formes fléchies, ainsi que de nombreuses données associées comme la fréquence, une représentation phonologique, un exemple de catégorie syntaxique possible etc. Seule la représentation orthographique nous intéresse. En effet, il est nécessaire de comparer des formes phonologiques homogènes tant au niveau du format que du processus de phonétisation, les représentations phonologiques proposées ne sont donc pas utilisées et les 130 000 formes que contient ce lexique sont phonétisées avec l'outil LIA_PHON.

7.3.1.2. ANALYSE

Sur 114 segments erronés traités, à savoir les 118 segments non reconnus moins les quatre « ɾ » non comptabilisés comme erreur, seuls 35 ont la même forme phonologique que la norme. Toutefois, ce nombre est à nuancer.

Nous avons évoqué précédemment (cf. 4.2.2. et 6.3.3.) le système phonologique sur lequel repose notre schéma d'annotation. Il s'agit d'un système restreint de 32 phonèmes alors que les systèmes ordinaires en dénombrent 36. Rappelons en effet que ce système restreint certaines oppositions vocaliques, comme :

- [e] / [ɛ], exemple « fé /fe/» (**1596**, *fait* /fɛ/)[6]
- [œ] / [ø] / [ə], exemple « plerer /pløre/ » (**1226**, *pleurer* /plœre/)
- [ɛ̃] / [œ̃], exemple « kopun /kopœ̃/ » (**1138**, *copains* /kopɛ̃/)

Ainsi, des erreurs, considérées dans les systèmes classiques comme des erreurs altérant la valeur phonique, seront considérées, par notre système, comme des erreurs de sélection orthographique (erreurs n'altérant pas la valeur phonique). 26 erreurs, que nous considérons comme étant des erreurs de sélection orthographique, ne sont ainsi pas reconnues par cette méthode de détection, en raison du système phonologique sur lequel est basé le module LIA_PHON. Si ces erreurs étaient comptabilisées, le nombre de segments que nous aurions à traiter serait de 61 segments. Néanmoins, les détecter à l'aide de LIA_PHON nécessiterait de modifier le système de règles ce qui demanderait un travail plus approfondi que nous ne pouvons pas mener dans le cadre de ce mémoire.

Notons également que certaines formes peuvent paraitre conformes phonologiquement, mais le contexte graphémique dans lequel est placé chaque graphème en

[6] Les représentations phonologiques mentionnées ici sont celles données automatiquement par le module LIA_PHON.

fait varier sa valeur. Par exemple, dans la production **1166**, l'élève ayant voulu écrire *petit* a écrit « petie ». Si l'on considère chaque graphème isolément, on obtient la forme phonologique / p - ə - t - i /, mais la présence du graphème muet « e » après la séquence *ti* va inciter le module LIA_PHON à transcrire le graphème *t* par sa valeur de position /s/. Le segment « petie » est transcrit phonétiquement /pəsi/.

Un petit nombre d'erreurs est dû à des problèmes d'accentuation (que les accents soient omis, insérés ou substitués à un autre accent), à l'exemple de « reveille » (**1296**, *réveille*) qui, en l'état, ne respecte pas la phonologie et ne peut donc être géré par notre méthode. Cependant, l'outil LIA_PHON contenant un module de réaccentuation, il pourrait être intéressant de faire un prétraitement à l'aide de ce module. Toutefois, ces erreurs vont souvent de pair avec d'autres erreurs qui, sans changer la phonologie, peuvent gêner la reconnaissance du mot par le module de réaccentuation, comme pour le segment « apré » (**1172**, *après*[7]). S'intéresser à ces cas nécessiterait une analyse du module spécifique.

Un certain nombre de formes respectent la phonologie mais ne seront pas traitées ici car elles contiennent des erreurs de segmentation et ne sont donc pas reconnues par TreeTagger comme segment unique, à l'exemple de « dincou » (**1226**, *d'un coup*), ou encore « à prè » (**1146**, *après*).

Enfin, les erreurs restantes sont dues à des erreurs de code phonographique, qu'elles soient agrémentées d'erreurs de segmentation ou non, ou sont des formes tronquées que TreeTagger ne peut donc analyser correctement.

Total			471
<unknown>			118
	Erronés		
		Phonologie respectée	**35**
		Phonologie respectée (dans système phonologique restreint)	26
		Accents non normés	2
		Phonologie altérée	51
	Logographes		4
Reconnus			**353**
	Mal reconnus (lemme)		42
		Avec erreur	41
	Bien reconnus		311
		Avec erreur	9
		Sans erreur	302

Tableau 20. Résumé des erreurs reconnues, étude sur un corpus restreint de 20 productions

[7] Production qui n'appartient pas au corpus restreint.

7.3.1.3. AUGMENTATION DU CORPUS DE TRAVAIL

Dans la suite de ce travail, nous nous concentrerons donc sur les 35 segments à phonologie normée. Ce nombre étant trop restreint, il nous a semblé pertinent de doubler notre corpus restreint afin de mieux en apprécier les finesses. Ce sous-corpus élargi contient désormais 40 productions totalisant 936 segments, dont 89 répondent à nos critères.

Cette augmentation assez conséquente des formes à traiter s'explique par le fait que les productions que nous avons ajoutées, pour des raisons dues au hasard, contiennent de nombreuses formes commençant par une majuscule comme « Il », « Bébé » ou encore « Chat » que TreeTagger ne reconnait pas malgré qu'elles soient bien orthographiées. Pour éviter que ces formes soient reconnues comme des erreurs alors qu'elles n'en sont pas, un prétraitement est nécessaire pour convertir ces majuscules en minuscules. Il reste alors 77 segments à traiter.

7.4. METTRE EN PLACE UNE MÉTHODE DE TRAITEMENT

Ce travail préliminaire avait uniquement pour but d'identifier les formes qu'il nous faudrait théoriquement traiter. La méthode employée était exclusivement manuelle et non automatisable puisqu'elle partait de la connaissance de la forme normée. En informatique, cette connaissance préalable que l'humain construit par analyse linguistique n'existe pas. Notre travail consiste donc à élaborer une méthode qui permette de trouver la forme normée correspondant à la forme écrite par l'élève.

Nous avons précédemment émis deux hypothèses, la première selon laquelle nous pourrons retrouver la forme normée par comparaison phonologique et la seconde selon laquelle nous pouvons, pour cela, nous appuyer sur le contexte du corpus. Pour pouvoir comparer phonologiquement les formes erronées à des formes normées possibles, il nous faut une liste de formes phonétiques ou à phonétiser. Nous pourrions prendre, à l'exemple des outils d'étiquetage morphosyntaxique, la liste de toutes les formes fléchies du français, toutefois un grand nombre de formes fléchies sont homophones en français et l'ambigüité entre ces formes serait trop importante. Il nous faut alors élaborer une pondération entre les homophones ou tout au moins une liste plus restreinte pour laquelle nous nous appuierons sur notre connaissance du contexte.

7.4.1. À PARTIR DU CORPUS

7.4.1.1. LEXIQUE ÉLABORÉ À PARTIR DU CORPUS

Comme nous l'a montré l'étude lexicométrique exposée en début de ce mémoire, certains éléments de lexique sont très présents dans notre corpus. Il s'agit d'éléments lexicaux fortement reliés au contexte de recueil du corpus (4 images). Ce constat nous amène à penser que nous pouvons nous appuyer sur celui-ci pour établir une liste de formes normées.

De plus, le nombre de productions dont nous disposons étant relativement important, nous pouvons faire l'hypothèse que la majorité des mots utilisés dans les productions se retrouvent dans plusieurs productions différentes sous des graphies différentes mais que la graphie majoritaire est la graphie normée. Ceci est une hypothèse très forte que nous allons devoir vérifier, mais si elle s'avérait exacte, nous pourrions imaginer comparer les différentes formes ayant même phonologie et normer les autres formes avec la forme la plus fréquente mais cela impliquerait que la forme la plus fréquente soit une forme normée. Afin de tester cette hypothèse, nous avons classé par fréquences les 1189 formes différentes que contiennent 258 productions[8]. Il en résulte que pour les formes phonologiques /ʃa/ (/chaa/ format LIA) et /mamɑ̃/ (/mmaamman/ format LIA), les formes les plus fréquentes sont *chat* et *maman* qui sont normées, néanmoins pour la forme phonologique /tɔ̃b/ (/ttonbb format LIA), la forme la plus fréquente est **tonbe*, forme non normée.

La forme majoritaire n'est donc pas toujours la forme normée. Nous faisons alors une nouvelle hypothèse selon laquelle au moins une des apparitions d'une forme est correctement orthographiée. Afin de mettre en application cette hypothèse, un lexique des formes normées de notre corpus est élaboré. Pour ce faire, toutes les productions ont été analysées par TreeTagger qui a produit, en sortie, une liste des segments du corpus étiquetés. À l'aide d'un script, les formes considérées comme erronées (étiquette <*unknown*>) sont supprimées, de même que les doublons. Ce traitement signifie que, comme précédemment, nous considérons comme normée toute forme ayant une étiquette différente de l'étiquette <*unknown*>. Après calcul de la représentation phonologique des segments conservés, on obtient un lexique des formes fléchies du français incluses dans notre corpus, soit 340 segments, et leurs représentations phonologiques.

[8] Nombre de productions transcrites au 19/03/2015.

Puis, pour chacun des 77 segments, une recherche est effectuée dans ce lexique afin de connaitre toutes les occurrences de leurs représentations phonologiques. Les résultats sont résumés dans le tableau ci-dessous :

Aucun correspondant trouvé	***18***
Un correspondant trouvé	***36***
Lemme attendu mais flexion attendue non trouvée	4
Lemme attendu non trouvé	2
Plusieurs correspondants trouvés	***23***
Forme attendue non trouvée	1
Total	**77**

Tableau 21. résumé des correspondants trouvés

Dans 18 cas, aucun équivalent n'a été trouvé, cela signifie que la forme concernée n'a jamais été correctement orthographiée, c'est le cas de la forme *dinosaure*, une seule occurrence, graphiée « dinosore » (**1166**), ou de la forme *chatte,* régulièrement graphiée « chate » (**810**, **2030**, **2908**…).

Parmi les formes ayant trouvé au moins un correspondant dans notre dictionnaire, 36 formes ont trouvé un correspondant unique. Celle-ci s'avère être la forme normée correspondante dans 30 cas. Pour les 6 cas restants, nous différencions les cas comme « fraire » (**586**, *frères*) reconnu *frère*, des cas comme « dor » (**812** et **1165**, *dort*) reconnus *dore*. Dans le premier cas, le lemme attendu est correctement reconnu, seule la marque du pluriel est absente. Une analyse syntaxique nous permettra de l'ajouter, nous pouvons donc considérer que la forme correspondante est trouvée. Dans le deuxième cas, en revanche, le lemme de la forme trouvée est DORER, tandis que le lemme de la forme attendue est DORMIR, il s'agit d'un problème plus délicat que nous ne pouvons pas juger comme acceptable.

Enfin, pour les 23 segments restants, plusieurs correspondants phoniques ont été trouvés, il sera donc nécessaire, dans un travail ultérieur, de désambigüiser ces formes en sélectionnant la forme normée correspondante. Pour l'heure, seul nous intéresse de savoir si la forme attendue a été trouvée ou non par notre méthode. C'est toujours le cas, sauf pour l'exemple « cou » (**1184**, *coup*) pour lequel seules les formes *cou* et *cous* ont été trouvées.

En résumé :

Formes attendues identifiées	56
Formes attendues non identifiées	21
Total	**77**

Tableau 22. Score de rappel pour le lexique issu du corpus

La méthode, qui consistait à établir un lexique à partir du corpus et qui se base sur l'hypothèse forte que tous les mots sont correctement orthographiés au moins une fois, permet donc déjà de repérer un certain nombre de formes attendues. Cependant, elle n'est pas encore satisfaisante puisqu'elle ne permet pas de traiter 22 cas sur 78 dont nous n'avons encore fait, pour la plupart, aucune analyse car ils ne correspondaient à aucune forme de notre lexique, il nous faut donc élargir celui-ci.

7.4.2. ÉLARGIR NOTRE LEXIQUE

Nous pouvons faire l'hypothèse que pour certains lemmes ayant de nombreuses formes associées, comme les verbes, même si toutes les formes utilisées ne sont pas normées, il existe dans notre corpus au moins une forme normée. Si cette hypothèse se vérifie, nous pouvons alors étendre notre lexique en déclinant toutes les formes des lemmes représentés dans notre lexique et espérer augmenter nos chances de reconnaissance des formes attendues. Prenons l'exemple de « dor » (**812** et **1165**, *dort*), la forme normée *dort* n'apparait pas en corpus, néanmoins ce verbe représente une action dessinée sur une des images, il est donc très présent dans le corpus sous des réalisations très diverses, nous pouvons donc supposer qu'au moins une de ses formes est normée.

Nous allons, pour vérifier cette hypothèse, élaborer à nouveau notre lexique en y ajoutant le champ lemme donné par TreeTagger. Pour chacun des lemmes présents dans notre corpus, toutes les formes associées à ce lemme sont recherchées dans la base lexicale **Lexique 3** et placées dans un nouveau dictionnaire phonétisé à l'aide de LIA_PHON. Notre dictionnaire, qui contenait 340 formes, contient désormais 1862 formes.

La même méthode que précédemment est appliquée à partir de ce dictionnaire élargi. Comme nous nous y attendions, ce nouveau dictionnaire permet de trouver les formes normées de 3 formes verbales supplémentaires, les formes « quogne » (**1166**, *cogne*), « pran » (**1666**, *prend*) et « vien » (**562**, *vient*). Il permet également d'associer, en plus des formes du verbe DORER, les formes du verbe DORMIR à la forme « dor » (**812** et **1165**, *dort*). En revanche, cette méthode ne permet pas de trouver la forme attendue de « cou » (**1184**, *coup*), ni des 16 formes restantes pour lesquelles aucun correspondant n'a été trouvé. On obtient donc pour le lexique simple issu du corpus et pour le lexique augmenté issu du corpus :

	Corpus simple	Corpus augmenté
Formes attendues identifiées	56	61
Formes attendues non identifiées	21	16
Total	**77**	**77**

Tableau 23. Comparaison du nombre de formes identifiées

Naturellement, cette méthode augmente également le nombre de formes normées trouvées et avec elles le degré d'ambigüité. Il est nécessaire de minimiser le nombre de formes correspondantes pour diminuer le degré d'ambigüité et pour faciliter les traitements ultérieurs et pour optimiser les chances de proposer une forme désambigüisée correcte. En comparant le nombre de formes trouvées, on obtient :

Nombre de correspondants trouvés	**Corpus simple**	**Corpus augmenté**
0	18	15
1	36	6
2	13	35
3	10	3
4		18
Total	**77**	

Tableau 24. Comparaison du degré d'ambigüité

Il nous faut donc trouver une méthode qui nous permette de trouver un maximum de formes attendues (réduire le silence) tout en minimisant le nombre de formes trouvées par forme attendue. Pour nous aider à comparer les différentes méthodes, nous utiliserons trois mesures :

- le **rappel**, le nombre de formes pour lesquelles la forme attendue a été trouvée ;

- le **degré d'ambigüité**, nombre moyen de formes trouvées à l'exclusion des cas où aucune forme n'a été trouvée ;

- la **précision**, le nombre de formes pour lesquelles la forme attendue a été trouvée parmi les formes ayant trouvé au moins une forme normée.

Mesure	**Corpus simple**	**Corpus augmenté**
Rappel	72.7%	79.2%
Degré d'ambigüité	1.6 forme	2.5 formes
Précision	94.9%	98.4%

Tableau 25. Comparaison des scores selon trois mesures des lexiques tirés du corpus

Précisons que les mesures du rappel et de la précision atteignent des scores inhabituellement élevées parce qu'elles portent sur des formes désambigüisées. Précisons également que si la précision n'atteint pas les 100%, c'est en partie dû à la présence, dans les lexiques issus du corpus, de formes attestées en langue française mais erronées en corpus. Le

premier lexique élaboré à partir du corpus contient ainsi 131 formes reconnues par TreeTagger, et donc considérées comme normées, mais lorsqu'on les replace en contexte dans le corpus, on s'aperçoit que ce sont des formes erronées. On citera par exemple le segment « chah » (**1594**, *chat*), reconnu par TreeTagger comme une forme fléchie du lemme CHAH et donc présent dans notre lexique. On dénombre ainsi dans ce lexique, sur un total de 340 entrées, 131 formes qui ne devraient pas y figurer car fausses en contexte.

7.4.2.1. LEXIQUES EXTERNES AU CORPUS

Nous venons de voir que l'utilisation du contexte restreint ne donne pas le résultat escompté, puisque le rappel n'excède pas les 80%. Nous allons donc devoir recourir à de nouveaux lexiques, extérieurs au corpus. Afin de faire un choix parmi les nombreux lexiques existants, nous nous sommes intéressé aux listes utilisées en milieu scolaire ainsi qu'aux listes employées par N. Catach (1995) pour étudier le système orthographique. Nous avons notamment recensé les listes lexicales suivantes :

– La **liste de fréquence lexicale de l'Éducation nationale** (disponible sur *Éduscol*[9]), cette liste contient 15 000 lemmes parmi les plus fréquents en langue française. Elle a été élaborée par E. Brunnet à partir de textes littéraires ou non.

– **Vocabulaire Orthographique de Base** (V.O.B., F. Ters, D. Mayer, G. Reichenbach, 1970), qui regroupe les 800 lemmes les plus fréquents dans les textes et classés par centres d'intérêts.

– **Gougenheim 2.00** (informatisée par B. New à partir des travaux de Gougenheim, 1964) qui contient 8 774 lemmes et qui a été élaboré à partir de productions orales ; seuls les mots les plus fréquents y figurent.

– **Manulex** (Lété, Sprenger-Charolles et Colé, 2004), base de données qui permet, à partir de l'étude de manuels scolaires, de « décrire la langue française écrite adressée à l'enfant en école primaire » à l'aide de listes de formes triées par fréquence et par niveau scolaire.

À l'exception des listes présentes dans **Manulex**, toutes ces listes sont constituées de lemmes. Pour pouvoir les utiliser, il nous faudrait donc les développer en formes fléchies de

[9] Portail national des professionnels de l'éducation. eduscol.education.fr

la même manière que nous l'avons fait pour le lexique élaboré à partir du corpus. Ce qui impliquerait, comme pour le lexique issu du corpus augmenté, d'y ajouter des formes peu employées par les scripteurs débutants.

Nous nous intéresserons donc uniquement à **Manulex**. Cet outil contient plusieurs bases selon le niveau des manuels scolaires étudiés, il permet de chercher au niveau CP, CE1, cycle 3 (CE2-CM2) ou sur tous les niveaux (CP-CM2). Sans représenter le contexte restreint du corpus, il s'apparente tout de même au contexte élargi de notre corpus, puisqu'il tire sa source d'écrits scolaires.

Il a pour but de représenter la langue française adressée à l'enfant et non celle que produit l'enfant, mais nous faisons l'hypothèse qu'il y a corrélation entre les deux. Afin de tester cette hypothèse, nous comparons les résultats donnés par différentes listes : **Manulex CP** puisqu'il s'agit du niveau scolaire de notre corpus, **Manulex CP-CM2** car nous avons rapidement fait le constat que certains jeunes scripteurs utilisent un vocabulaire très large à l'exemple de « rquonsili » (**1280**, *réconcilie*) mais également **Lexique 3** considéré comme liste représentative du lexique d'un scripteur adulte hors contexte contenant toutes les formes fléchies de la langue française au moment de son élaboration, il nous permettra donc de vérifier l'influence du contexte, notre hypothèse de départ.

Les listes de mots données par **Manulex** contiennent à la fois les formes fléchies mais aussi leurs catégories associées ainsi que différentes mesures de fréquence. Il peut donc y avoir, contrairement à **Lexique 3**, plusieurs fois la même forme si le lemme d'origine est différent. Afin de pouvoir comparer ces listes aux listes tirées de notre corpus et à **Lexique 3**, les formes homonymes ne seront considérées qu'une seule fois, indépendamment de leur catégorie.

Comme précédemment, après phonétisation des listes, nous pouvons mesurer leur efficacité :

	Corpus simple	Corpus augmenté	Manulex CP	Manulex CP-CM2	Lexique 3
Formes attendues identifiées	56	61	74	77	77
Formes attendues non identifiées	21	16	3	0	0
Total	**77**				

Tableau 26. Comparaison du nombre de formes identifiées

Nombre de correspondants trouvés	Corpus simple	Corpus augmenté	Manulex CP	Manulex CP-CM2	Lexique 3
0	18	15	3	0	0
1	36	6	15	1	1
2	13	35	29	24	23
3	10	3	18	29	17
4		18	8	11	11
5			4	7	9
6				4	14
7				0	0
8				1	2
Total	77				

Tableau 27. Comparaison du degré d'ambigüité

Mesure	Corpus simple	Corpus augmenté	Manulex CP	Manulex CP-CM2	Lexique 3
Rappel	72.7%	79.2%	96.1%	100%	100%
Degré d'ambigüité	1.6 forme	2.5 formes	2.4	3.2	3.7
Précision	94.9%	98.4%	100%	100%	100%

Tableau 28. Comparaison des scores selon trois mesures

7.5. COMBINER LES LISTES

Ce tableau nous permet d'énoncer plusieurs constats. En premier lieu, il apparait que la liste **Manulex CP-CM2** et **Lexique 3** présentent des taux de précision et de rappel des formes identifiées, mais Lexique 3 fait montre d'un degré d'ambigüité plus important. Ce constat confirme qu'il est plus avantageux d'utiliser une liste restreinte tirée du contexte scolaire qu'un lexique regroupant toutes les formes fléchies d'un locuteur adulte. Cependant, il n'existe pas de réelle fracture entre la liste tirée du contexte scolaire **Manulex CP–CM2** et la liste générale **Lexique 3**. Ceci peut s'expliquer par une augmentation de l'ambigüité dans **Manulex** due à une prise en compte des noms propres. Or, on constate que les enfants en début d'école primaire utilisent très peu de noms propres et lorsqu'ils en utilisent, à l'exemple de « Ronron » (**1284**, *Ronron*), ceux-ci ne sont pas répertoriés dans la base **Manulex**. Nous pourrions donc envisager d'abaisser le degré d'ambigüité en ne prenant pas en compte les noms propres.

On constate également une fracture relativement nette entre les listes tirées du corpus (**Corpus simple** et **Corpus augmenté**) et les listes scolaires et générales (**Manulex CP**, **Manulex CP-CM2** et **Lexique 3**). On observe ainsi une augmentation des performances au niveau du rappel, mais également au niveau de la précision. Néanmoins, l'augmentation de

ces performances a pour conséquence l'augmentation du degré d'ambigüité. Au vu de ce constat, il serait intéressant d'utiliser plusieurs listes afin d'éviter une trop grande augmentation du degré d'ambigüité. Ainsi, on appliquerait un système de priorité entre les lexiques. Seuls les segments non reconnus par le lexique le plus restreint et présentant le moins d'ambigüité seraient analysés par le deuxième lexique, le degré d'ambigüité s'en trouvera réduit.

Enfin, ce tableau nous montre également très clairement, qu'il est plus intéressant d'utiliser la liste **Manulex CP** que le lexique augmenté tiré du corpus (**Corpus augmenté**). En effet, la précision et le rappel y sont plus importants tandis que le degré d'ambigüité est le même voire moindre. Ce constat peut s'expliquer par la méthode employée pour élaborer ce lexique qui consistait à décliner toutes les formes à partir des lemmes présents, ce qui implique pour les verbes par exemple de nombreuses formes peu pertinentes, comme le subjonctif imparfait *tombât*, peu voire jamais usité par les enfants.

Afin d'éviter ces formes peu pertinentes nous pouvons imaginer développer le lexique tiré du corpus non pas à l'aide de **Lexique 3** mais à l'aide de **Manulex CP**. Cette méthode devrait présenter deux avantages. Elle devrait permettre de développer le lexique dans la limite des formes les plus utilisées au CP, les formes comme *tombât* ne sont pas présentes dans **Manulex CP**. Mais elle devrait permettre également de ne pas conserver certaines erreurs *chah* (reconnues par TreeTagger et donc conservées dans les deux lexiques tirés du corpus) qui est une forme inexistante dans **Manulex CP**.

Comme précédemment avec **Lexique 3**, la liste **Manulex CP** est envoyée à TreeTagger, afin d'obtenir les lemmes de chaque forme. On reprend l'extraction des lemmes du lexique tiré du corpus réalisée précédemment avec TreeTagger également. Chaque fois qu'une forme est associée à un lemme présent dans la liste **Manulex CP** et également présent dans notre extraction de lemmes, la forme est ajoutée dans un nouveau lexique. À partir de ce nouveau lexique, on obtient :

Mesure	**Corpus simple**	**Corpus augmenté**	**Corpus augmenté 2** avec Manulex CP	**Manulex CP**
Rappel	72.7%	79.2%	80.5%	96.1%
Degré d'ambigüité	1.6 forme	2.5 formes	2.1	2.4
Précision	94.9%	98.4%	98.4%	100%

Tableau 29. Comparaison du nouveau lexique augmenté aux autres lexiques

Sans présenter les avantages de la liste **Manulex CP**, ce lexique présente tout de même un degré d'ambigüité moindre. Dans notre optique d'améliorer les résultats en combinant les listes, nous pouvons envisager de combiner trois listes : le dernier lexique tiré du corpus et **Manulex CP**, qui présentent toutes deux un degré d'ambigüité relativement bas associé à des scores de précision et rappel plutôt élevés, et **Manulex CP-CM2** dont l'avantage est de présenter un taux de précision de 100%. Les résultats obtenus sont donnés ci-dessous :

Mesure	**Corpus augmenté 2 + Manulex CP-CM2**	**Manulex CP + Manulex CP-CM2**
Rappel	98.7%	100%
Degré d'ambigüité	1.6	2.4
Précision	98.7%	100%

Tableau 30. Résultats des méthodes par combinaison de listes

Chacune de ces méthodes présente un avantage certain. En effet, utiliser la liste **Manulex CP** permet d'atteindre des taux de précision et de rappel de 100%, tout en limitant le degré d'ambigüité, tandis qu'utiliser un lexique tiré du corpus permet de restreindre très fortement l'ambigüité. En revanche, cela revient également à accepter un certain degré de bruit, c'est-à-dire de formes ayant trouvé des correspondants mais dont aucun ne correspond à la forme attendue.

Or, nous ne faisons pour l'instant que rechercher les formes normées possibles. Plus tard viendra la phase de désambigüisation, qui s'effectuera à partir des formes normées que nous trouvons avec notre méthode. Accepter un taux de rappel inférieur à 100%, c'est accepter de ne pas présenter toutes les formes normées attendues lors de la désambigüisation et donc accepter que certains segments ne seront pas désambigüisés correctement et donc non reconnus correctement. Afin de permettre le meilleur taux de réussite possible, nous privilégierons un taux de précision et de rappel de 100% plutôt qu'un faible degré d'ambigüité. Nous nous appuierons donc sur **Manulex CP** et **Manulex CP-CM2**.

7.6. CONCLUSION

Comme nous l'avions émis en hypothèse, le contexte peut nous aider à reconnaitre les formes normées. Néanmoins, nous ne pouvons utiliser ni un contexte trop proche, car contenant trop d'erreurs et pas assez vaste, ni le contexte de la langue en général, car ouvrant la porte à trop d'ambigüités.

Afin d'être conscient des limites de la méthode que nous venons de présenter, il est également nécessaire d'évaluer le nombre de segments qui, sans être visés par cette méthode seront également pris en compte. Il s'agit de segments en corpus qui ne correspondent pas à des segments à phonologie normée mais pour lesquels des formes normées seront trouvées. Le plus souvent, ces formes normées trouvées ne correspondent pas à la forme attendue. Par exemple, notre méthode va prendre en compte le segment « dar » (**3006**, *dort*) et va lui faire correspondre la forme normée *dard*. Ceci s'explique par sa proximité phonologique plus importante à la forme *dard* qu'à la forme *dort*. En se limitant au corpus restreint de 10 productions, 21 des 83 segments restants seront ainsi reconnus, de manière erronée, par notre méthode. Ces formes sont à considérer comme du bruit, on obtient donc une augmentation du bruit produit par notre méthode.

7.6.1. ALGORITHME D'ANNOTATION DES ERREURS

À partir des conclusions précédentes, un algorithme d'annotation des erreurs a pu être élaboré. Il permet la recherche de correspondants ainsi que l'annotation des erreurs, à partir de ces correspondants. Les méthodes employées pour l'annotation des erreurs sont encore expérimentales et fonctionnent pour l'instant sur un corpus restreint de trois segments « cha » (**1986**, *chat*) qui présente une omission de lettre muette, « tonba » (**1956**, *tomba*) qui présente une substitution orthographique et « soire » (**3066**, *soir*) présentant une insertion de lettres muettes.

L'annotation automatique se déroule en plusieurs étapes. En premier lieu, il s'agit d'étiqueter le corpus à l'aide de TreeTagger pour pouvoir repérer les erreurs à traiter. Rappelons que nous ne traitons que les formes étiquetées <*unknown*> par TreeTagger. Ces formes extraites peuvent alors être phonétisées à l'aide de LIA_PHON. Puis, leur représentation phonologique est comparée aux représentations phonologiques de toutes les formes contenues dans la liste **Manulex CP**. Si aucun correspondant n'est trouvé, la même opération est répétée avec les formes de la liste **Manulex CP-CM2**. Chaque fois qu'une forme correspond et qu'elle n'a pas déjà été trouvée, les graphies du couple segment en corpus / forme correspondante sont comparées.

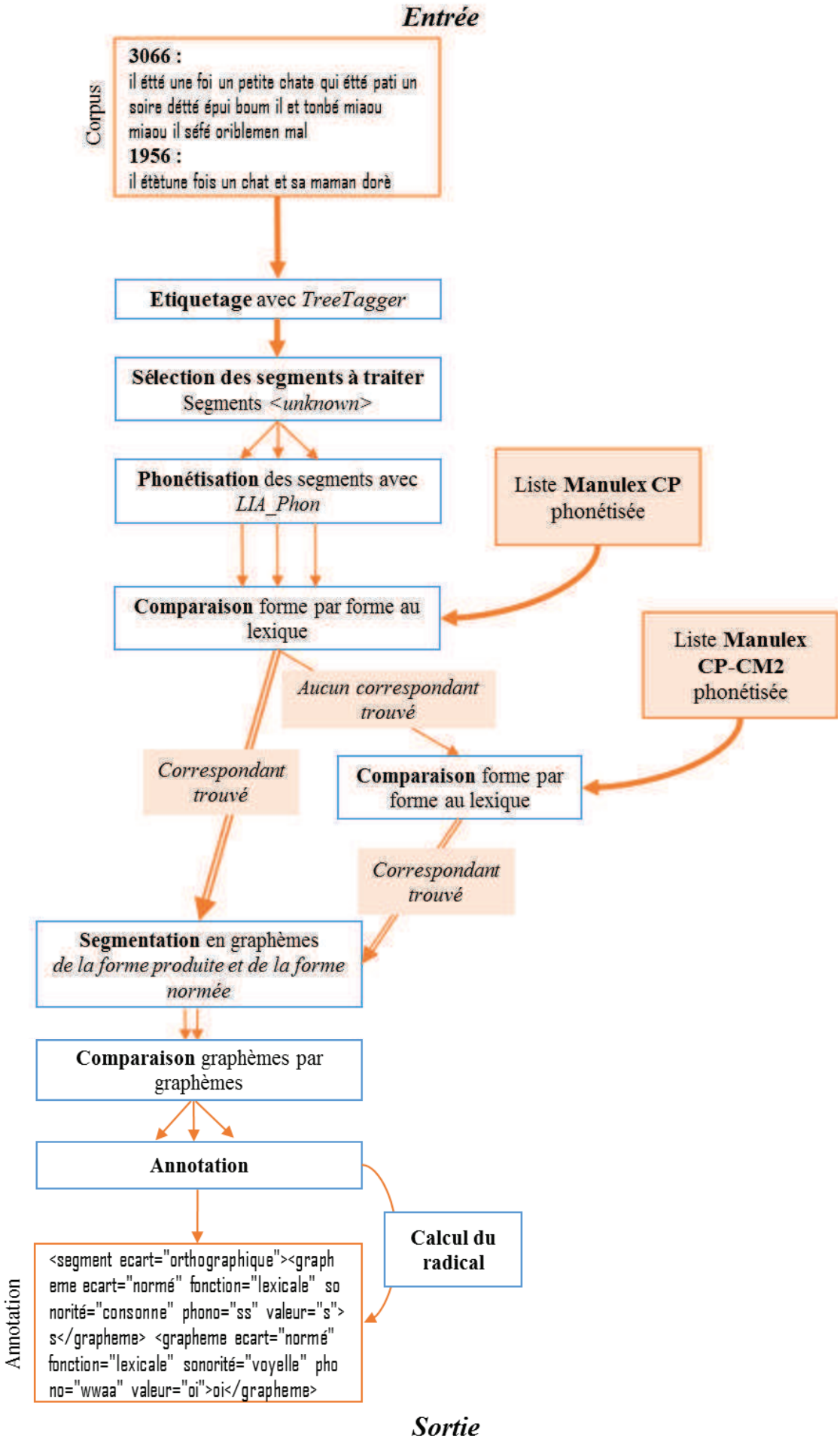

Entrée
Corpus
3066 :
il étté une foi un petite chate qui étté pati un soire détté épui boum il et tonbé miaou miaou il séfé oriblemen mal
1956 :
il étètune fois un chat et sa maman dorè
Etiquetage avec TreeTagger
Sélection des segments à traiter
Segments <unknown>
Phonétisation des segments avec LIA_Phon
Liste Manulex CP phonétisée
Comparaison forme par forme au lexique
Aucun correspondant trouvé
Liste Manulex CP-CM2 phonétisée
Correspondant trouvé
Comparaison forme par forme au lexique
Correspondant trouvé
Segmentation en graphèmes de la forme produite et de la forme normée
Comparaison graphèmes par graphèmes
Annotation
Calcul du radical
Annotation
<graph eme ecart="normé" fonction="lexicale" so norité="consonne" phono="ss" valeur="s"> s</grapheme> <grapheme ecart="normé" fonction="lexicale" sonorité="voyelle" pho no="wwaa" valeur="oi">oi</grapheme>
Sortie

Figure 7. Structure de l'algorithme d'annotation des erreurs de sélection orthographique

Pour faciliter la comparaison, les deux formes sont transformées en séquence de graphèmes et en séquence de phonèmes à l'aide d'un module perl élaboré spécialement et appelé *graph.pm* (annexe 11). Il permet de segmenter en graphèmes une forme et de faire correspondre chaque graphème à un phonème ou à une absence de phonème. Puis, les graphèmes sont comparés un à un et sont annotés selon leur écart à la norme : normé, omis, inséré ou substitué orthographiquement. Ne nous intéressant qu'aux erreurs à phonologie normée, nous ne traitons pas les autres cas. Pour chaque couple de graphèmes sont également spécifiées :

- la sonorité, selon que le graphème soit une consonne, une voyelle, une semi-consonne ou un graphème muet ;
- la valeur phonique grâce à l'association établie plus haut ;
- la valeur graphique de la forme normée ;
- la valeur graphique de la forme attestée en corpus.

Enfin, la fonction, grammaticale ou lexicale, du graphème est également donnée. Pour la déterminer, un deuxième module a été élaboré. Il permet à partir d'une forme normée et de sa catégorie de retrouver le radical. Si le graphème concerné se trouve dans le radical, la fonction lexicale lui est attribuée, sinon on dira qu'il s'agit d'un graphème grammatical.

La sortie de ce programme est un tableau contenant la forme en corpus, la forme normée, le lemme, la catégorie syntaxique et l'annotation (cf. tableau).

Forme en corpus	**Forme normée**	**Lemme**	**Catégorie**
cha	chats	chat	NOM
Annotation : <grapheme ecart="normé" fonction="lexicale" sonorite="consonne" phono="ch">ch</grapheme> <grapheme ecart="normé" fonction="lexicale" sonorite="voyelle" phono="aa">a</grapheme> <grapheme ecart="omis" fonction="lexicale" sonorite="muet" valeur="t"></grapheme> <grapheme ecart="omis" fonction="grammaticale" sonorite="muet" valeur="s"></grapheme> 			
cha	chat	chat	NOM
 <grapheme ecart="normé" fonction="lexicale" sonorite="consonne" phono="ch">ch</grapheme> <grapheme ecart="normé" fonction="lexicale" sonorite="voyelle" phono="aa">a</grapheme> <grapheme ecart="omis" fonction="lexicale" sonorite="muet" valeur="t"></grapheme> 			
soire	soir	soir	NOM
 <grapheme ecart="normé" fonction="lexicale" sonorite="consonne" phono="ss">s</grapheme> <grapheme ecart="normé" fonction="lexicale" sonorite="voyelle" phono="wwaa">oi</grapheme> <grapheme ecart="normé" fonction="lexicale" sonorite="consonne" phono="rr">r</grapheme> <grapheme ecart="inséré" fonction="lexicale" sonorite="muet">e</grapheme></grapheme> 			

soire	soirs	soir	NOM
 <grapheme ecart="normé" fonction="lexicale" sonorite="consonne" phono="ss">s</grapheme> <grapheme ecart="normé" fonction="lexicale" sonorite="voyelle" phono="wwaa">oi</grapheme> <grapheme ecart="normé" fonction="lexicale" sonorite="consonne" phono="rr">r</grapheme> <grapheme ecart="orthographique" fonction="grammaticale" sonorite="muet" valeur="s">e</grapheme> 			
tonba	tomba	tomber	VER:simp
 <grapheme ecart="normé" fonction="lexicale" sonorite="consonne" phono="tt">t</grapheme> <grapheme ecart="orthographique" fonction="lexicale" sonorite="voyelle" phono="on" valeur="om">on</grapheme> <grapheme ecart="normé" fonction="lexicale" sonorite="consonne" phono="bb">b</grapheme> <grapheme ecart="normé" fonction="grammaticale" sonorite="voyelle" phono="aa">a</grapheme> 			

Tableau 31. Exemple de sortie du traitement des erreurs de sélection orthographique

Actuellement, ce programme ne permet que de traiter les segments « cha », « tonba » et « soire ». Pour l'étendre à d'autres segments, il suffit d'étendre les règles permettant la segmentation en graphèmes contenues dans le script graph.pm (annexe 11) et d'étoffer les règles permettant le calcul des radicaux dans *radical.pm*. En effet, à ce jour, seuls les radicaux des noms et des verbes peuvent être calculés. En revanche, il n'est nul besoin de modifier le programme principal.

7.7. CONCLUSION

La présente étude a permis de montrer que comparer les représentations phonologiques des formes erronées et des formes issues de lexiques du français s'avérait une méthode que l'on pouvait juger efficace pour traiter les erreurs de sélection orthographique, ce qui constituait notre première hypothèse de travail.

Cette étude montre également qu'il n'est pas possible, comme nous le pensions au début, de partir du corpus pour trouver les formes normées attendues sans s'accorder une marge d'erreur. Néanmoins, l'utilisation d'un contexte de production plus large reste possible puisque nous utilisons des listes issues du contexte scolaire, tout comme notre corpus.

Enfin, cette étude valide, au moins en partie, notre schéma d'annotation, puisque nous montrons qu'il est possible à partir d'outils automatiques d'annoter certaines erreurs de notre corpus selon les termes prévus par notre schéma.

CHAPITRE 8 - CONCLUSION ET PERSPECTIVES

Le schéma d'annotation que nous avons élaboré doit désormais être mis à l'épreuve de la pratique en élargissant les traitements et le corpus testé. Il sera donc certainement amené à évoluer. Pour l'heure, nous avons montré dans le chapitre précédent (cf. chapitre 7), à partir d'une ébauche de traitement, qu'il permettait d'annoter une partie des erreurs de sélection orthographique. Le travail effectué ici n'est donc qu'un travail préliminaire qui doit permettre de nombreux traitements supplémentaires.

8.1. À PARTIR DE LEXIQUES

Tout comme nous l'avons fait pour les erreurs traitées au chapitre 7, certaines erreurs phonographiques et orthographiques peuvent être annotées à l'aide de lexiques de formes phonétiques. Un traitement préliminaire sera cependant nécessaire.

8.1.1. TRAITEMENT DES ERREURS ORTHOGRAPHIQUES PAR PHONOLOGIE ÉTENDUE

Rappelons que nous avons considéré comme erreurs de sélection orthographique les erreurs comme « kopun » (**1138**, *copains*). Ces erreurs peuvent être considérées comme altérant la valeur phonique si l'on se place dans le système phonique classique du français à 36 phonèmes, mais comme sans conséquence sur la valeur phonique dans un système restreint à 31 phonèmes (cf. 6.3.3.).

Si ces erreurs ne sont pas traitées actuellement par notre méthode, c'est uniquement parce que nous utilisons un système de phonétisation (LIA_PHON) basé sur un système phonétique classique. Afin de nous conformer aux choix faits lors de l'élaboration de notre schéma, il nous faudrait donc modifier les règles du système en remplaçant les phonèmes /e/ et /ɛ/ par l'archiphonème /E/, les phonèmes /o/ et /ɔ/ par l'archiphonème /O/ et les phonèmes /ø/, /œ/ et /ə/ par l'archiphonème /œ/. Nous pourrons alors calculer à nouveau les formes phonétiques de chaque forme des lexiques et des formes erronées pour appliquer les mêmes traitements que pour les erreurs à phonologie normée, encore appelées erreurs de sélection orthographique (cf. chapitre 7). Notons que cet élargissement des règles aura certainement un impact sur le nombre de correspondants que l'on peut associer à une forme et donc au degré d'ambigüité.

8.1.2. TRAITEMENT DES ERREURS DE VALEUR

Les erreurs de valeur sont des erreurs issues d'une méconnaissance des règles d'interaction entre les graphèmes, à l'exemple de « kə » (**1064**, *que*). Pour traiter ces erreurs, il est possible, lors du calcul de la valeur phonique de la forme erronée, d'attribuer à chaque graphème toutes les correspondances phonographiques possibles. Ainsi, en considérant que le graphème *c* peut avoir pour valeur phonique /k/ et /s/ et que le graphème *e* peut avoir pour valeur phonique /E/, /œ/ ou être muet, on obtient les possibilités suivantes : /sE, sœ, s, kE, kœ, k/. Chaque représentation fait alors l'objet d'une comparaison aux formes phonétisées des lexiques, comme pour le traitement des erreurs précédentes.

8.1.3. TRAITEMENT DES ERREURS DE SUBSTITUTION FLEXIONNELLE

Ces erreurs ne sont pas très nombreuses dans notre corpus et sont souvent impliquées dans la chaine sonore, par exemple « fesa » (**1168**, *fit*). Mais elles sont très liées à la morphologie et leur nombre devrait donc augmenter dans les classes supérieures. Elles correspondent, dans les classifications présentées dans le chapitre 3, aux erreurs de génération. Nous pouvons reprendre la méthodologie élaborée pour traiter ces erreurs. Pour rappel, il s'agissait d'extraire le radical de la forme erronée d'une part et les informations flexionnelles d'autre part pour proposer une forme normée fléchie selon ces informations. Dans le cas où le radical comporterait des erreurs de sélection orthographique, une comparaison avec un lexique de radicaux phonétisés pourrait être envisagée.

8.1.4. TRAITEMENT DES ERREURS RÉCURRENTES

Pour traiter les erreurs les plus fréquentes, nous pouvons essayer de reprendre la métaphore de la traduction automatique proposée pour les SMS (cf. 3.3.3.). Celle-ci proposait de considérer certaines erreurs récurrentes comme une autre norme, à l'exemple de **tjs* et **tjrs* (*toujours*). Comme nous l'avons vu au chapitre 2 (cf. 2.4.1.), notre corpus comprend quelques erreurs récurrentes, à l'exemple de la forme *tombe* plus fréquemment écrite sous la forme **tonbe* que sous sa forme normée. Mais beaucoup d'erreurs ne sont pas aussi systématiques.

De même qu'un certain nombre de substitutions utilisées dans les SMS peuvent s'expliquer par un format d'écriture consonantique (comme la substitution de **tjs* ou **tjrs* à

toujours, ou de *svp* à *s'il-vous-plait*), nous pouvons également donner des principes généralisables à un certain nombre d'erreurs contenues dans notre corpus. Par exemple, nous avons montré dans ce même chapitre 2, qu'un grand nombre d'erreurs correspondaient à une omission de lettres muettes ou à une substitution de graphèmes à même valeur phonique. Nous pourrions alors décliner notre lexique, de manière à obtenir, selon ces observations, des formes erronées plausibles associées de leur forme normée.

8.1.5. PREMIERS TRAITEMENTS DES ERREURS DE CODE PHONOGRAPHIQUE

Après ces différentes corrections, il est également possible d'essayer de corriger quelques erreurs de code phonographique à l'aide de méthodes basées sur des comparaisons graphiques à l'aide de lexiques, telles que les distances d'édition (Levenshtein, 1966). Cependant, il est fort probable que peu d'erreurs puissent être corrigées ainsi. En effet, les erreurs de code phonographique s'accompagnent généralement d'erreurs de sélection orthographique, à l'exemple du segment « donbre » (**1953**, *tombe*). Nous pouvons tout de même espérer détecter les inversions de graphèmes comme « apèr » (**3006**, *après*).

8.2. ANALYSES SYNTAXIQUES

Il nous semble difficile d'envisager les autres erreurs présentées dans notre schéma, sans passer par une analyse syntaxique de notre corpus. Afin d'optimiser cette analyse, il est préférable de repérer les unités syntaxiques dans un premier temps. Comme nous l'avons vu au chapitre 2, plusieurs indices sont à notre disposition pour cela. Nous pouvons notamment envisager de nous baser sur les points et les virgules lorsqu'ils ne sont pas adjacents à un retour à la ligne contraint par la fin de la ligne (symbolisé par un slash simple / dans les transcriptions). Cette restriction nous permettra de ne pas envisager les cas où le point marque le changement de ligne et non la phrase. Nous pourrons également nous appuyer sur les retours à la ligne volontaires (symbolisés par un slash double // dans les transcriptions) et sur les connecteurs comme *et*, *et après*, *puis* et *ensuite*.

L'analyse syntaxique est complexifiée par la présence de nombreux segments erronés mais devrait être facilitée par la relative simplicité de la plupart des phrases produites par les jeunes scripteurs. Elle devrait nous permettre de traiter les erreurs de segmentation mais également de sélectionner la forme attendue parmi les formes normées données aux étapes précédentes.

8.2.1. ENVISAGER LA SEGMENTATION

Comme nous l'avons mentionné au chapitre 4, P. Cappeau et M.-N. Roubaud (2005) ont étudié les erreurs de segmentation et en ont proposé quelques tendances générales. Notamment le fait que les formes agglutinées en un segment unique correspondent souvent à des mots grammaticaux monosyllabiques, comme *à*, *se*, *y*, etc., comme « yavè » (**1166**, *y avait*), ou encore « tonbil » (**1666**, *tombe, il*). Ils ont également observé que les redécoupages effectués par les enfants correspondent souvent à des mots également existants, à l'exemple de « à prè » (**1156**, *après*) ou encore « des sendu » (**1336**, *descendu*).

Ces observations pourront donc nous aider à faire des hypothèses de segmentation qui devront être validées ou invalidées par une analyse syntaxique. La plupart des productions présentant peu d'erreurs de code phonographique (erreurs altérant la valeur phonique), il pourrait également être intéressant de tester certaines méthodes de segmentation utilisées en reconnaissance vocale, à partir des formes phonétisées des productions.

8.2.2. DÉSAMBIGÜISER LES FORMES NORMÉES

La plupart des traitements que nous venons de décrire s'appuient sur des lexiques, ce qui nous donne des résultats englobant différentes formes normées possibles. Il faut alors désambigüiser ces ensembles afin de trouver la forme normée attendue.

8.2.2.1. DÉSAMBIGÜISER LES FIGURES DE MOTS

Les figures de mots correspondent à des formes normées, c'est-à-dire présentes en lexique, mais qui ne sont pas les formes attendues en contexte. Seule une analyse syntaxique permet de détecter ces erreurs et de les corriger. N. Catach (1995) a proposé une liste des logogrammes les plus fréquents. En nous appuyant sur cette liste, nous pourrions établir les différentes formes normées possibles, puis une analyse syntaxique pourrait nous aider à les désambigüiser en se basant principalement sur les catégories grammaticales. En effet, la plupart des oppositions logogrammiques présentes dans nos corpus confrontent des mots ou groupes de mots de catégories différents (a_{VERBE} / $à_{PRÉPOSITION}$, est_{VERBE} / $et_{CONJONCTION}$, $c'est_{PRONOM+VERBE}$ / $ses_{DÉTERMINANT}$, etc.). Les oppositions homophoniques *est* / *et* observées dans le sous-corpus de 40 productions utilisé au chapitre 7 ont été reportées dans le tableau suivant (tableau 32), seules les erreurs reconnues comme telles par la méthode utilisant TreeTagger sont répertoriées :

	CONTEXTE GAUCHE	HOMOPHONE	CONTEXTE DROIT	FORME NORMEE ATTENDUE
SEGMENT	**partêre**	**é**	**il**	**et**
CATEGORIE	VER:simp	VER:pper	PRO:PER	
SEGMENT	**mal**	**é**	**il**	**et**
CATEGORIE	ADV	VER:cond	PRO:PER	
SEGMENT	**pleur**	**é**	**il**	**et**
CATEGORIE	NOM	ADJ	PRO:PER	
SEGMENT	**tombe**	**ét**	**aprè**	**et**
CATEGORIE	VER:pres	VER:pper	NOM	
SEGMENT	**elle**	**er**	**tomber**	**est**
CATEGORIE	PRO:PER	VER:futu	VER:infi	

*Tableau 32. Oppositions homophoniques (*et/est*) et leur contexte*

À travers ces observations et les observations faites tout au long de notre travail, il apparait que les formes *est* / *et* peuvent très souvent être désambigüisés à l'aide d'indices simples. En effet, le connecteur *et* est généralement placé devant le groupe sujet, constitué d'un pronom ou d'un groupe nominal. Il peut également, comme c'est le cas dans notre tableau, être placé devant un autre connecteur comme *après* ou *ensuite* (« é an site », **1558**, *et ensuite*). Dans de très rares cas, il est utilisé à l'intérieur d'un groupe nominal, comme « Sa maman et sai fraire » (**586**, *Sa maman et ses frères*). La forme *est*, quant à elle, est très souvent utilisée comme auxiliaire. Elle se trouve donc généralement entre un nom ou un pronom et un participe passé. Ces deux formes apparaissent donc dans des contextes très distincts qui pourront nous aider à les désambigüiser.

8.2.2.2. DÉSAMBIGÜISATION PAR FRÉQUENCE

Les premières analyses réalisées sur les erreurs de sélection orthographique à partir d'un corpus de 40 productions (cf. chapitre 7) tendent à montrer qu'une désambigüisation par fréquence pourrait donner des résultats tout à fait acceptables, exceptions faites des figures de mot, pour lesquels d'autres moyens sont disponibles.

Nous avons répertorié les fréquences de chaque forme normée proposée par la méthode élaborée au chapitre 7, les fréquences étant données par les listes **Manulex** et **Lexique 3**. Puis, nous avons sélectionné la forme ayant la fréquence la plus élevée (tableau 33). Nous n'avons pas considéré les cas où les différentes formes normées relèvent toutes du même lemme. Il s'agit d'ambigüité flexionnelle et non lexicale, ces cas pourront être désambigüisés à l'aide d'analyses syntaxiques.

Représentation phonologique	**Forme normée attendue**	*Manulex CP*	*Manulex CP-CM2*	*Lexique 3*
ddii	dit	dit (VER)	dit (VER)	dit
ddoorr	dort	dort	dort	dort
ffai	fait	fait (VER)	fait (VER)	fait
kkan	quand	quand	quand	quand
kkou	coup	coup	coup	coup
kkrrii	cri	crie (VER)	crie (VER)	cri
ppllaass	place	place (NC)	place (NC)	place
pprran	prend	prend	prend	prend
ppuyii	puis	puis	puis	puis
ssee	se	se	se	se
ttonbb	tombe	tombe (VER)	tombe (VER)	tombe
ttou	tout	tout (ADV)	tout (ADV)	tout

Tableau 33. Résultats d'un tri par fréquence

Les listes Manulex étant à la fois classées par formes et par catégorie, dans le cas de formes homonymes, la catégorie de la forme la plus fréquente est inscrite entre parenthèses. Sont surlignées en orange les formes ne correspondant pas à la forme normée. Elles sont très peu nombreuses, ce qui indique que l'utilisation de fréquence peut sans doute aider à la désambigüisation des formes normées trouvées à l'aide des lexiques. Une analyse morphologique et syntaxique est ensuite nécessaire pour déterminer la flexion.

Les méthodes avancées ici ne sont que des pistes de réflexion qui nécessitent d'être mises en pratique afin de vérifier leur validité. De plus, elles ne suffiront pas à traiter toutes les erreurs. D'autres méthodes devront alors être mises au point.

BIBLIOGRAPHIE

Abney, S. (1991). Parsing by chunks. *Principle-Based Parsing*. Kluwer Academic Publishers, pages 257–278.

D'Alessandro, C. (2014), Synthèse de la parole, *L'information grammaticale*, mars 2014, 141, 31-36.

Antionadis G., Echinard S., Kraif O., Lebarbé T., Ponton, C., & Echinard, S. (2005). Modélisation de l'intégration de ressources TAL pour l'apprentissage des langues : la plateforme MIRTO. *Revue ALSIC. Apprentissage des Langues et Systèmes d'Information et de Communication*, *8*(2).

Aubergé, V. (1985). Passage automatique du texte orthographique vers le texte phonétique. *14e JEP-GALF*.

Aw A., Zhang M., Xiao J. & Su J. (2006). A phrase-based statistical model for sms text normalization. *Proceedings of the COLING/ACL on Main conference poster sessions,* 33-40.

Baranes. M, (2012). Vers la correction automatique de textes bruités: Architecture générale et détermination de la langue d'un mot inconnu. *RECITAL'2012 - rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, Grenoble, France, 95-108.

Baranes, M. & Sagot, B. (2014). Normalisation de textes par analogie: le cas des mots inconnus. *TALN - Traitement Automatique du Langage Naturel*, Marseille, France. 137-148.

Beaufort, R, Roekhaut, S., Cougnon, L.-A. & Fairon, C. (2010). Une approche hybride traduction/correction pour la normalisation des SMS, *Actes de la 17e conférence sur le traitement automatique des langues naturelles (TALN'10)*, Montréal.

Béchet, F. (2001). LIA_PHON - Un système complet de phonétisation de textes, *Traitement Automatique des Langues (T.A.L.*) 42, 1/2001, Hermès.

Bégin, C., Saint-Laurent, L. & Giasson J. (2005). La contribution des écritures provisoires dans la réussite en orthographe : étude longitudinale. *Revue Canadienne de Linguistique Appliquée / CJAL. 8.2,* 148-166.

Belrhali, R. (1995) : Phonétisation automatique d'un lexique général du français : systématique et émergence linguistique, (thèse de doctorat, Université Stendhal, Grenoble, France).

Benzitoun, C., Fort, K., & Sagot, B. (2012, June). TCOF-POS: un corpus libre de français parlé annoté en morphosyntaxe. In JEP-TALN 2012-Journées d'Études sur la Parole et conférence annuelle du Traitement Automatique des Langues Naturelles (pp. 99-112).

Blanchard, A., Kraif, O. & Ponton, C. (2009). Mastering Overdetection and Underdetection in *Learner-Answer Processing: Simple Techniques for Analysis and Diagnosis*. Université Stendhal Grenoble 3.

Blanche-Benveniste, C. & Chervel, A. (1978, éd. augmentée). L'orthographe. Paris : Maspero.

Blanche-Benveniste, C. & Chervel, A. (1969). L'orthographe. Paris : Maspero.

Boissière, P., Bouraoui, J. L., Vella, F., Lagarrigue, A., Mojahid, M., Vigouroux, N., & Nespoulous, J. L. (2007). Méthodologie d'annotation des erreurs en production écrite. Principes et résultats préliminaires. *Actes de TALN*, 2, 529-538.

Bouillon, P. (Ed.). (1998). Traitement automatique des langues naturelles. De Boeck Supérieur.

Boyd, A. (2009). Pronunciation modeling in spelling correction for writers of English as a foreign language. *Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume : Student Research Workshop and Doctoral Consortium*, Boulder, Colorado, 31–36.

Brill, E. & Moore, R. C. (2000). An improved error model for noisy channel spelling correction. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics,* Hong Kong, 286-293.

Cappeau, P. & Roubaud M.-N. (2005). *Enseigner les outils de la langue avec les productions d'élèves,* Paris : Bordas.

Catach, N. (1984). La phonétisation automatique du français: les ambigüités de la langue écrite. Paris : Editions du CNRS.

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

Catach, N., Duprez D. & Legris M. (1980). L'enseignement de l'orthographe : l'alphabet phonétique international, la typologie des fautes, la typologie des exercices, Paris : Nathan.

Catach, N. (1979), Le graphème, *Pratiques*(25), 21-32.

Catach, N., (1978, 10e édition, 2011). L'orthographe, *Que sais-je ?*, Paris : Presses Universitaires de France.

Chanier T., (1992), Perspectives de l'apport de l'EIAO dans l'apprentissage des langues étrangères: modélisation de l'apprenant et diagnostic d'erreurs. , M. 3 (4). 25-34.

Chanier T., (1996), Learning a Second Language for Specific purposes, within a hypermedia framework. *Computer-Assisted Language Learning (CALL),* 9(1), 3-43.

Chanquoy, L., & Negro, I. (2004). *Psychologie du développement*. Paris : Hachette Éducation.

Choudhury M., Saraf R., Jain V., Mukherjee A., Sarkar 1 S. & Basu A. (2007). Investigation and modeling of the structure of texting language. *International Journal on Document Analysis and Recognition,* 10 (3), 157–174.

Cook P. & Stevenson S. (2009). An Unsupervised Model for Text Message Normalization. *Proceedings of the workshop on Computational Approaches to Linguistic Creativity*, 71–78.

David, J. & Fraquet, S. (2011). L'écriture en action et actions de l'écriture à l'école maternelle, *Le Français Aujourd'hui*, 174.

David, J. (2006). L'orthographe du français et son apprentissage, historique et perspectives. *Orthographe en questions (L')*, 169-124.

David, J. (2003). Linguistique génétique et acquisition de l'écriture, *Faits de langue*, 22, 37-45.

Déchaux, C. & Milan, V. (2013). Les écritures approchées : les bénéfices de cet exercice. Mémoire Education.

Dietterich, T. G., Hild, H., & Bakiri, G. (1995). A comparison of ID3 and backpropagation for english text-to-speech mapping. *Machine Learning*, *18*(1), 51-80.

Ducard, D., Honvault, R., & Jaffré, J. P. (1995). L'orthographe en trois dimensions. Paris : Nathan.

Elalouf, M.-L. (2005). Ecrire entre 10 et 14 ans. CRDP de Versailles.

Fabre-Cols, C.(dir.) (2000) Apprendre à lire des textes d'enfants, *Savoirs en pratique*, Bruxelles : De Boeck Duculot.

Fayol, M. & Jaffré, J.-P. (2014). L'orthographe, *Que sais-je ?* , Paris : Presses Universitaires de France.

Fayol, M. (2013). L'acquisition de l'écrit*, Que sais-je* ?, Paris : Presses Universitaires de France.

Fayol, M., & Jaffré, J. P. (2008). Orthographier. Paris : Presses Universitaires de France.

Fayol, M. (1989). Une approche psycholinguistique de la ponctuation. Etude en production et en compréhension. *Langue française*. 81, 21-39.

Ferreiro, E. (2008). L'écriture avant la lettre. Paris : Hachette Éducation.

Ferreiro, E. et Gomez Palaccio, M. (1988). Lire-écrire à l'école, comment s'y prennent-ils ?, Paris : Presses Universitaires de France.

Fijalkow, J., Cussac-Pomel, J., & Hannouz, D. (2009). L'écriture inventée: empirisme, constructivisme, socioconstructivisme. *Éducation et didactique*(3), 63-97.

Fijalkow, J., & Liva, A. (1993). Clarté cognitive et entrée dans l'écrit. Jaffré, Sprenger-Chraolles et Fayol, Lecture-Ecriture : Acquisition, *Les Actes de la Villette*. Paris : Nathan Pédagogie, 203-229.

Fraquet, S., & David, J. (2013). Ecrire en maternelle : comment approcher le système écrit ? *Repères*, 47, 19-40.

Frith, U. (1985). Beneath the surface of developmental dyslexia. K. E. Patterson, JC Marshall, & M. Coltheart (Eds.), Surface dyslex-ia: Neuropsychological and cognitive studies of phonological reading, 301-330.

Fu, G. & Luke, K.-K. (2003). A two-stage statistical word segmentation system for Chinese. *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing -Volume 17*, Sapporo, Japan, 156–159.

Fuchs, C., Lacheret-Dujour, A., Victorri, B., Danlos, L., & Luzzati, D. (1993). Linguistique et Traitement automatiques des Langues. Hachette université langue, linguistique, communication.

Geoffre, T. (2013). Vers le contrôle orthographique au cycle 3 de l'école primaire: étude psycholinguistique et propositions didactiques (thèse de doctorat, université Stendhal-Grenoble, 3).

Gougenheim, G. (1964). L'élaboration du français fondamental (ler degré): étude sur l'établissement d'un vocabulaire et d'une grammaire de base (Vol. 1). Didier.

Ghneim, N. (1997). Relations entre les codes de l'oral et de l'écrit : contraintes et ambigüités, (thèse de doctorat, Université Stendhal-Grenoble 3).

Goigoux, R., & Cèbe, S. (2006). Apprendre à lire à l'école: tout ce qu'il faut savoir pour accompagner l'enfant. Retz.

Granger, S., Vandeventer, A. & Hamel, M.-J. (1998). Analyse des corpus d'apprenants pour l'ELAO basé sur le TAL. *Traitement automatique des langues* 42(2), Paris: Hermès, 609-621.

Grefenstette, G., & Tapanainen, P. (1994). What is a word, what is a sentence? Problems of tokenization. *3rd Conference on Computational Lexicography and Text Re-search*, 79-87.

Guimard, P. (2003). L'analyse clinique de l'orthographe lexicale chez l'enfant débutant ou en difficulté: De quelques repères théoriques et méthodologiques. *Glossa*, (84), 24-35.

Guimier De Neef, E. & Fessard, S. (2007). Evaluation d'un système de transcription de SMS. *Proceedings of the 26th International Conference on Lexis and Grammar,* 217–224.

Haton, J.-P., Cerisara, C., Fohr, D., Laprie, Y. & Smaïli, K. (2006). Reconnaissance automatique de la parole - Du signal à son interprétation, Paris : Dunod.

Heift, T. & Schulze, M. (2007). Errors and intelligence in computer-assisted language learning: Parsers and pedagogues. Oxom : Routledge, Taylor &Francis group.

Jacquet-Pfau, C. (2001). Correcteurs orthographiques et grammaticaux. *Revue française de linguistique appliquée*, 6(2), 81-94.

Jaffré, J. P. (2005). L'orthographe du français, une exception?. *Le français aujourd'hui*, (1), 23-31.

Jaffré, J. P., Bousquet, S., & Massonnet, J. (1999). Retour sur les orthographes inventées. Les dossiers des sciences de l'éducation : Des enfants, des livres et des mots, 1.

Jaffré, J. P. (1997). Des écritures aux orthographes: fonctions et limites de la notion de système. *Des orthographes et leur acquisition*, 19-36.

Jiang, F., Liu, H., Chen, Y., & Lu, R. (2004). An Enhanced Model for Chinese Word Segmentation and Part-Of-Speech Tagging. *Proceedings of the Third SIGHAN Workshop on Chinese Language* Processing, Barcelona, Spain, 28–32.

Kernighan, M.D., Church, K.W. & Gale, W.A. (1990). A spelling correction program based on a noisy channel model. In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, 205-210.

Kobus C., Yvon F. & Damnati G. (2008). Normalizing SMS: are two metaphors better than one? *Proceedings of the 22nd International Conference on Computational Linguistics*(1), Manchester, UK, 441–448.

Kraif, O., & Ponton, C. (2007). Du bruit, du silence et des ambigüités: que faire du TAL pour l'apprentissage des langues. *Actes de TALN*, Toulouse.

Kraif, O., (2005), Evaluation automatique de productions lexicales : une analyse à 4 niveaux. *UNTELE'2005*, Compiègne.

Kraif, O., Antionadis, G., Echinard, S., Lebarbé, T., Loiseau, M., & Ponton C., (2004), NLP Tools for CALL: the Simpler, the Better. *STIL/ICALL Symposium 2004. NLP and Speech Technologies in Advanced Language Learning Systems*, Italie, Venise.

Laporte, E., & Silberztein, M. (1989). Vérification et correction orthographiques assistées par ordinateur. *Actes de la Convention IA*. 89, 252.

Lebart, L. & Salem, A. (1994). Statistique Textuelle. Paris: Dunod.

Lété, B., Sprenger-Charolles, L., & Colé, P. (2004). Manulex: A grade-level lexical database from French elementary-school readers. *Behavior Research Methods, Instruments, & Computers, 36*, 156-166.

Léty, M. (1980). Transcription orthographique-phonétique: un système interpréteur (thèse de doctorat, Université Joseph-Fourier-Grenoble I).

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8), 707–710.

L'Haire, S. (2011). Traitement Automatique des Langues et Apprentissage des langues Assisté par Ordinateur : bilan, résultats et perspectives. (Thèse de doctorat, Université de Genève, Faculté des Lettres, Genève).

L'Haire, S. & Vandeventer Faltin A. (2003). Diagnostic d'erreurs dans le projet FreeText. *Apprentissage des Langues et Systèmes d'Information et de Communication*, 6 (2), 21-37.

Lucci V. & Millet A. (1994). L'orthographe de tous les jours. Enquête sur les pratiques orthographiques des Français, Paris : Champion.

Lurçat, L. (1985). L'écriture et le langage écrit de l'enfant en écoles maternelle et primaire, Paris : les éditions ESF.

Ma, W.-Y. & Chen, K.-J. (2003). Introduction to CKIP Chinese word segmentation system for the first international Chinese Word Segmentation Bakeoff. *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*, Sapporo, Japan, 168–171.

Mangu, L. & Brill, E. (1997). Automatic rule acquisition for spelling correction. *Proc. 14th International Conference on Machine Learning* (97), 187-194.

Menézo. J. (1999). CELINE, vers un correcteur lexico-syntaxique adaptatif et semi-automatique. (Thèse de doctorat, Institut National Polytechnique de Grenoble – INPG).

Mikheev, A. (2003). Text segmentation, *The Oxford Handbook of Computational Linguistics*, Oxford University Press, 201-218.

Mikheev, A. (2002). Periods, Capitalized Words, etc. *Computational Linguistics* 28 (3). 289–318.

Mikheev, A. (2000). Tagging Sentence Boundaries. *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, Seattle : WA, 264-271.

Mitton, R. (1996) English spelling and the computer. London : Longman.

Morel, M. & Lacheret-Dujour, A. (2001). « Kali : synthèse vocale à partir du texte. De la conception à la mise en œuvre ». *Traitement Automatique des Langues*, 42 (1), 193-221.

Ndiaye, M. et Vandeventer Faltin, A. (2004). Correcteur orthographique adapté à l'apprentissage du français. BULAG, 29:117–134.

New, B., Brysbaert, M., Veronis, J. & Pallier, C.(2007). The use of film subtitles to estimate word frequencies, In *Applied Psycholinguistics*, 28, 661-677

New, B. (2006). Lexique 3 : Une nouvelle base de données lexicales. *Actes de la Conférence Traitement Automatique des Langues Naturelles* (TALN 2006), avril 2006, Louvain, Belgique.

New, B., Pallier, C., Ferrand, L. & Matos R. (2001). Une base de données lexicales du français contemporain sur internet : LEXIQUE, *L'Année Psychologique*, 101, 447-462.

Ortéga, É., & Lété, B. (2010). *eManulex: Electronic version of Manulex and Manulex-infra databases*.

Pacton, S., Foulin, J. N., & Fayol, M. (2005). L'apprentissage de l'orthographe lexicale. *Rééducation orthophonique*, 43(222), 47-68.

Pacton, S., Fayol, M., & Perruchet, P. (2002). The acquisition of untaught orthographic regularities in French. *Precursors of functional literacy*, 121-137.

Palmer, D. D. (2000). Tokenisation and sentence segmentation. *Handbook of natural language processing*, 11-35.

Palmer, D. D. & Hearst, M.A. (1997). Adaptive Multilingual Sentence Boundary Disambiguation. *Computation Linguistics*, 23 (2), 241-267.

Park, Y. A., & Levy, R. (2011). Automated whole sentence grammar correction using a noisy channel model. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*(1), 934-944.

Passerault J.-M. (1991). La ponctuation, recherche en psychologie du langage. *Pratiques* (70), 85-104.

Pellat, J. C. (1988). Indépendance ou interaction de l'écrit et de l'oral? Recensement critique des définitions du graphème, CNRS, 133-146.

Pérennou G. (1991). Les vérificateurs orthographiques. *Texte et ordinateur. Les Mutations du Lire-Ecrire*, Actes du Colloque interdisciplinaire, Université Paris X Nanterre. 55-86.

Polguère, A. (2003). Lexicologie et sémantique lexicale: notions fondamentales. Montréal : Presses de l'Université de Montréal.

Reynar, J.C. & Ratnaparkhi, A. (1997). A Maximum Entropy Approach to Identifying Sentence Boundaries. *Proceedings of the fifth conference on Applied natural language processing*, Washington, D.C, 16-19.

Riley, M. D. (1989). Statistical tree-based modeling of phonetic segment durations. *The Journal of the Acoustical Society of America*, 85 (1), 44-44.

Romary L., Salmon-Alt S., Francopoulo G. (2004). Standards going concrete: from LMF to Morphalou. *Workshop on Electronic Dictionaries*, Coling 2004, Geneva, Switzerland.

Schmid, H. (2007). Tokenizing. *An International Handbook*. Corpus Linguistics, Berlin.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, 44–49.

Sejnowski, T. J., & Rosenberg, C. R. (1987). Parallel networks that learn to pronounce English text. *Complex systems*, *1*(1), 145-168.

Silberztein, M. (1993). Dictionnaires électroniques et analyse automatique de textes: le système INTEX. Masson.

Strube De Lima, V. L. (1990). Contribution à l'étude du traitement des erreurs au niveau lexico-syntaxique dans un texte écrit en français, (Thèse de doctorat, Université Joseph Fourrier, Grenoble 1).

Suignard, P. & Kerroua, S. (2013). Utilisation de contextes pour la correction automatique ou semi-automatique de réclamations clients. *Actes de TALN'13*, Les Sables d'Olonne, France, 699.

Ters, F., Mayer, G. & Reichenbach D. (1970). Programme de vocabulaire orthographique de base, Langue Française, 6, 125-126.

Torzec N., Moudenc T. & Emerard F. (2001). Prétraitement et analyse linguistique dans le système de synthèse tts cvox : Application à la vocalisation automatique d'e-mails. *Traitement Automatique des Langues* 42 (1), pp.17-46.

Toutanova, K. & Moore, R. C. (2002). Pronunciation Modeling for Improved Spelling Correction. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Philadelphie, États-Unis, 144–151.

Véronis, J. (1988). Morphosyntactic Correction in Natural Language Interfaces. *Proceedings, 13th International Conference on Computational Linguistics*, 708-713. International Committee on Computational Linguistics.

Vienney, S. & Bioud, M. (Ed.) (2004). Correction automatique : Bilan et perspectives, Bulag, 29, Centre Tesnière, Presse Universitaire de Franche-Comté.

Whitelaw, C., Hutchinson, B., Chung, G. Y., & Ellis, G. (2009). Using the web for language independent spellchecking and autocorrection. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*(2), 890-899.

Williams, B., & Maier, F. (1991). A Spelling Corrector for Use in Text-to-Speech Synthesis for English. In *Second European Conference on Speech Communication and Technology*.

Wu, A. (2003). Chinese word segmentation in MSR-NLP. *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing-Volume 17*, Sapporo, Japan, 172–175.

Xue, N. & Shen, L. (2003). Chinese word segmentation as lmr tagging. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing-Volume 17*, Sapporo, Japan, 176–179, Sapporo, Japan.

SITOGRAPHIE

Ivanic, R. & McEnery, T., 1996, The Lancaster Corpus of Children's Project Writing (LCCPW), http://www.lancaster.ac.uk/.

Marchand, P. (2008). Richesse lexicale. http://pascal-marchand.fr/spip.php?article13

New, B., Pallier, C., Ferrand, L. & Matos R. (2001). http://www.lexique.org

Ortéga, É., & Lété, B. (2010). http://www.manulex.org

Romary L., Salmon-Alt S., Francopoulo G. (2004). www.cnrtl.fr/lexiques/morphalou/ (ATILF/ Nancy Université - CNRS).

Véronis, J. (2007). Texte : Richesse lexicale. http://blog.veronis.fr/2007/03/texte-richesse-lexicale.html.

TABLE DES TABLEAUX

TABLE DES FIGURES

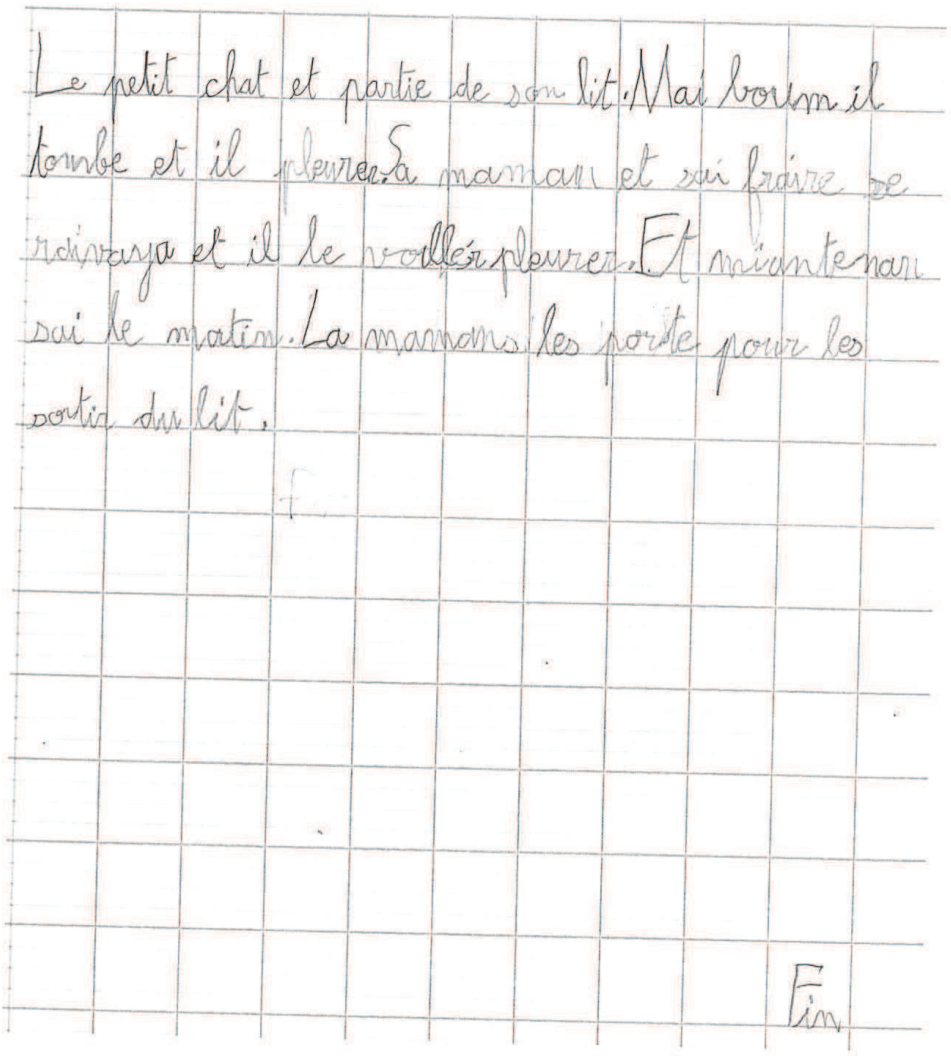
TABLE DES ANNEXES

ANNEXES

Annexe 1 :

Extrait de corpus

Productions écrites – CP


Le petit chat et partie de son lit. Mai bourm il tombe et il pleure. Sa maman et sa frère le rejouja et il le voulé pleurer. Et maintenant le matin. La mamans les porte pour les sortir du lit.

Fin

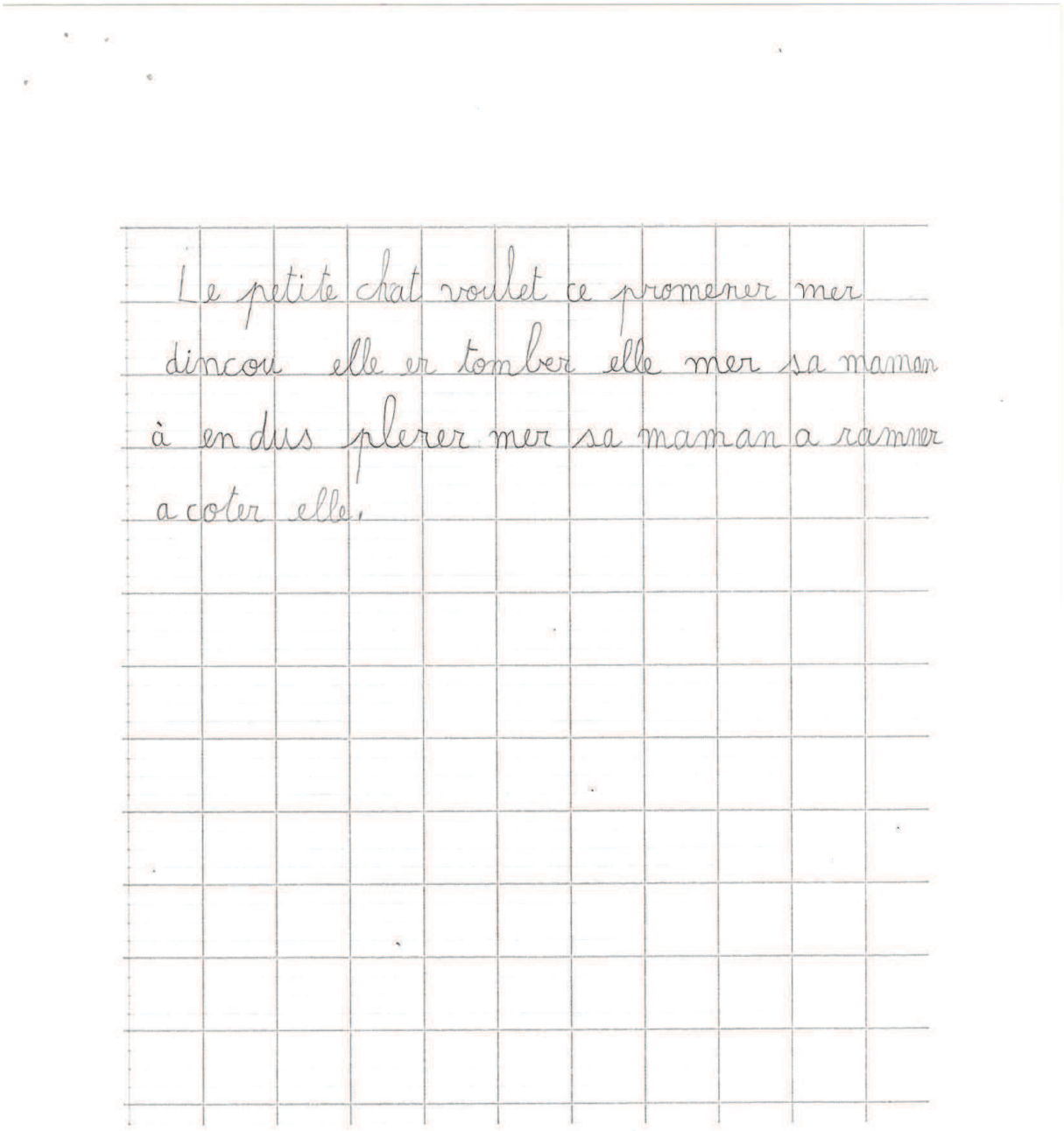
586

il été libéran des sotra à prè
il tombe et après sa fait
boum et après il fait misor deux
foi et sa mère la trape

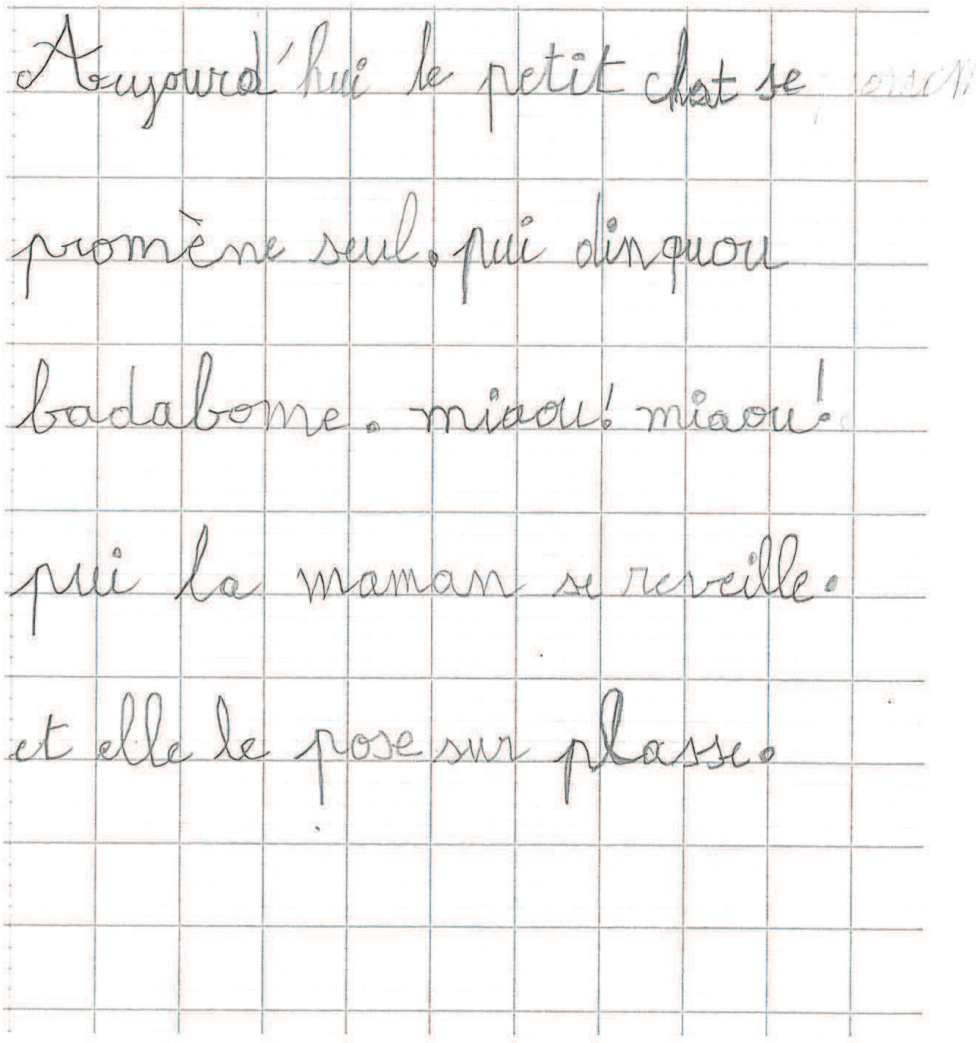
1156

Il y avè un petie chat qui entandai le qwi d'un
dinosore et ie se dwogne la tête et il plere et il
va chez sa maman.

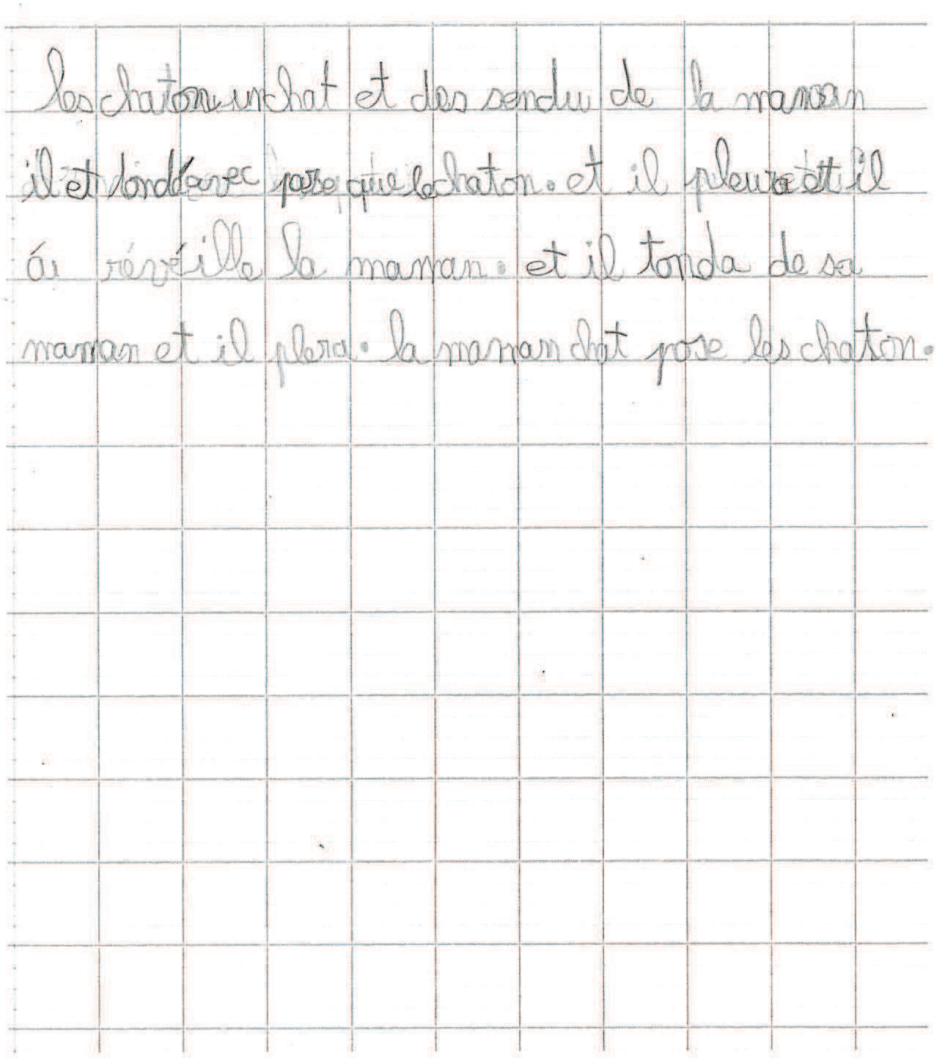
1166


Le petite chat voulet ce promener mer
dincou elle er tomber elle mer sa maman
à en dus plerer mer sa maman a rammer
a coter elle.

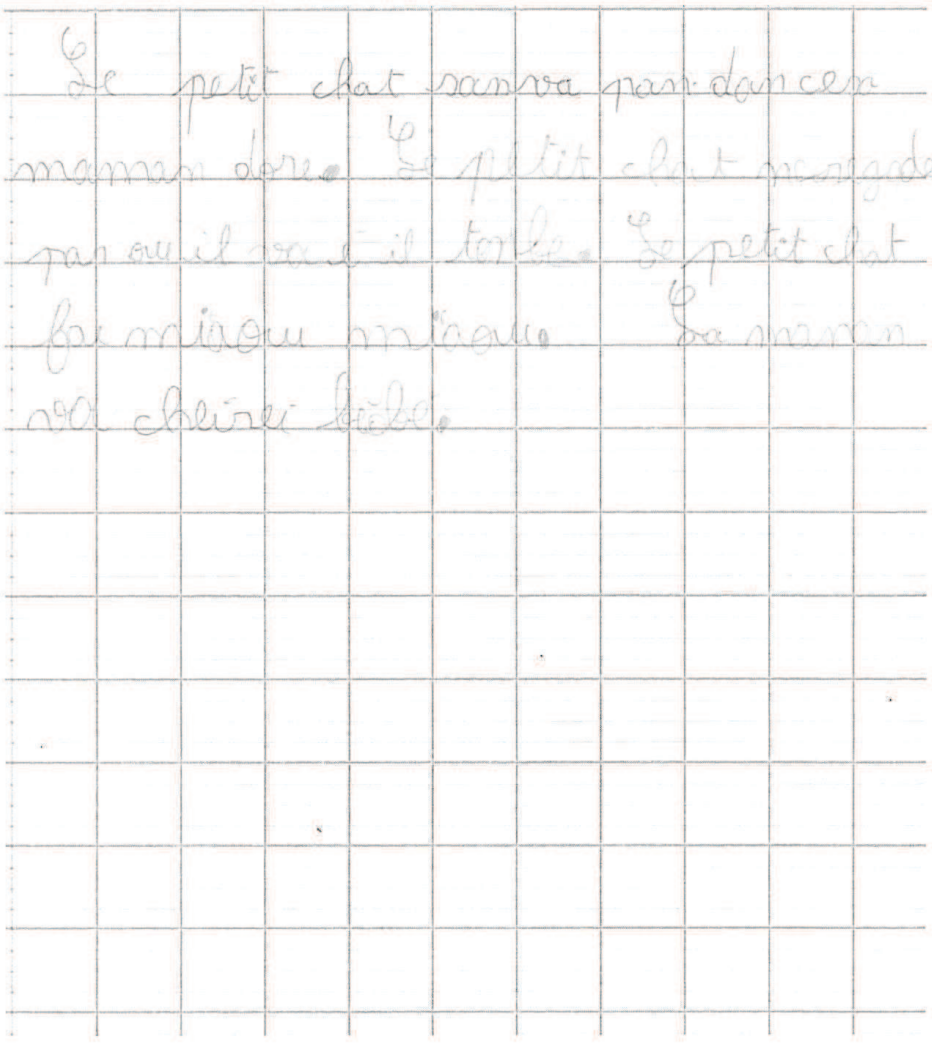
1276


Aujourd'hui le petit chat se promène seul. Puis dinquou badabome. miaou! miaou! puis la maman se reveille. et elle le pose sur plasse.

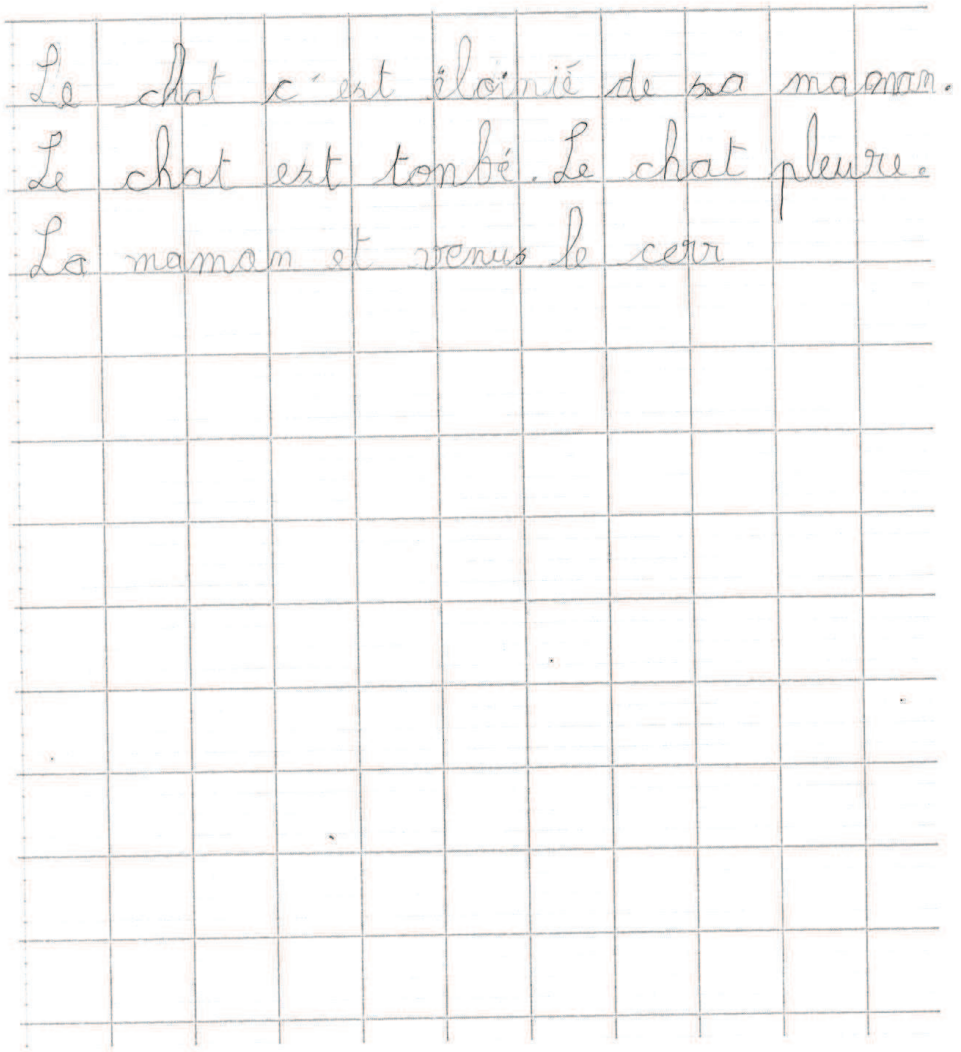
1296


les chaton un chat et des sendu de la maman
il et tondé avec parce que le chaton. et il pleure et il
á réveillé la maman. et il tonda de sa
maman et il plera. la maman chat pose les chaton.

1336


Le petit chat sauva par dans ces
maman dore. Le petit chat mange de
par ou il va il tombe. Le petit chat
fait miaou miaou. Sa maman
est chiree triste.

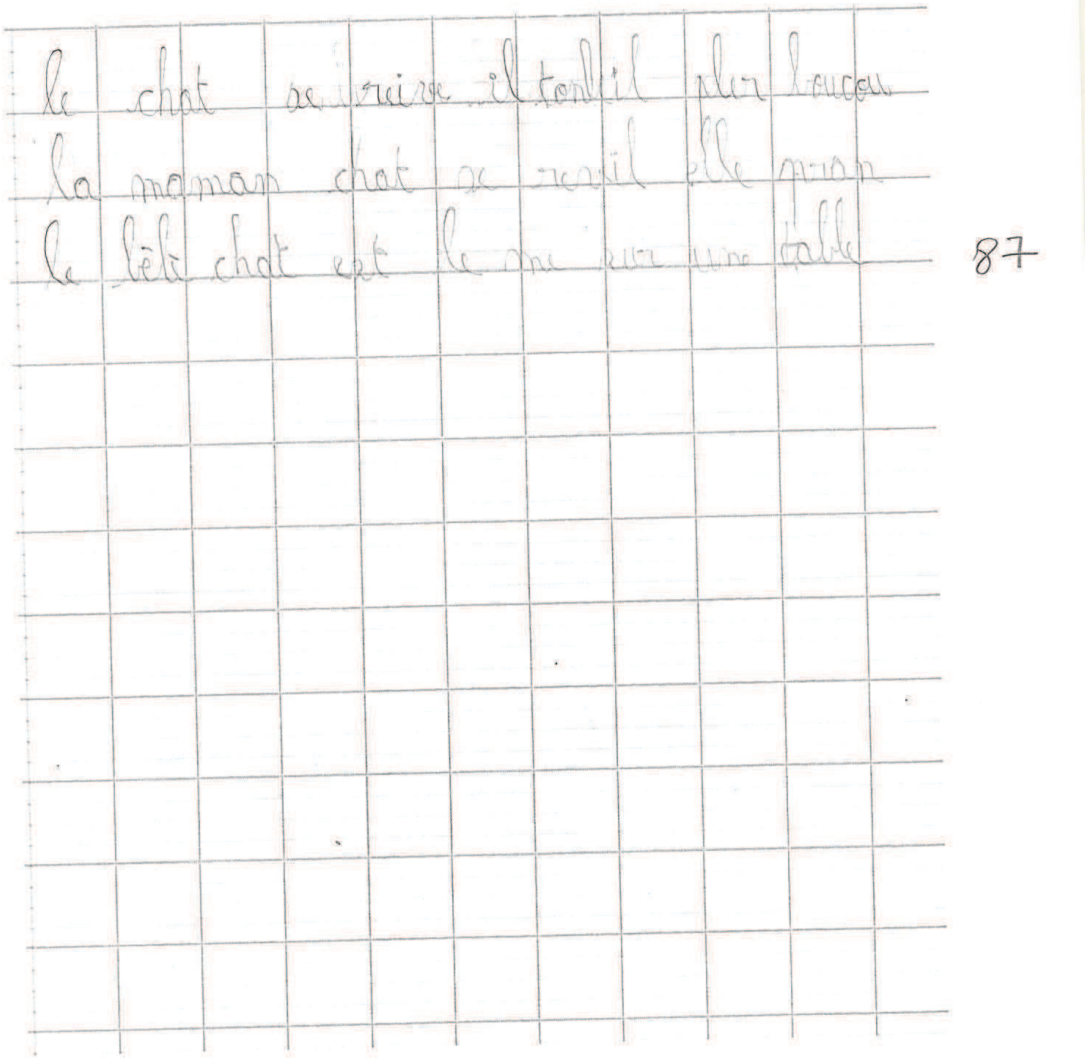
1346


Le chat c'est éloigné de sa maman.
Le chat est tombé. Le chat pleure.
La maman et venus le cerr

1556

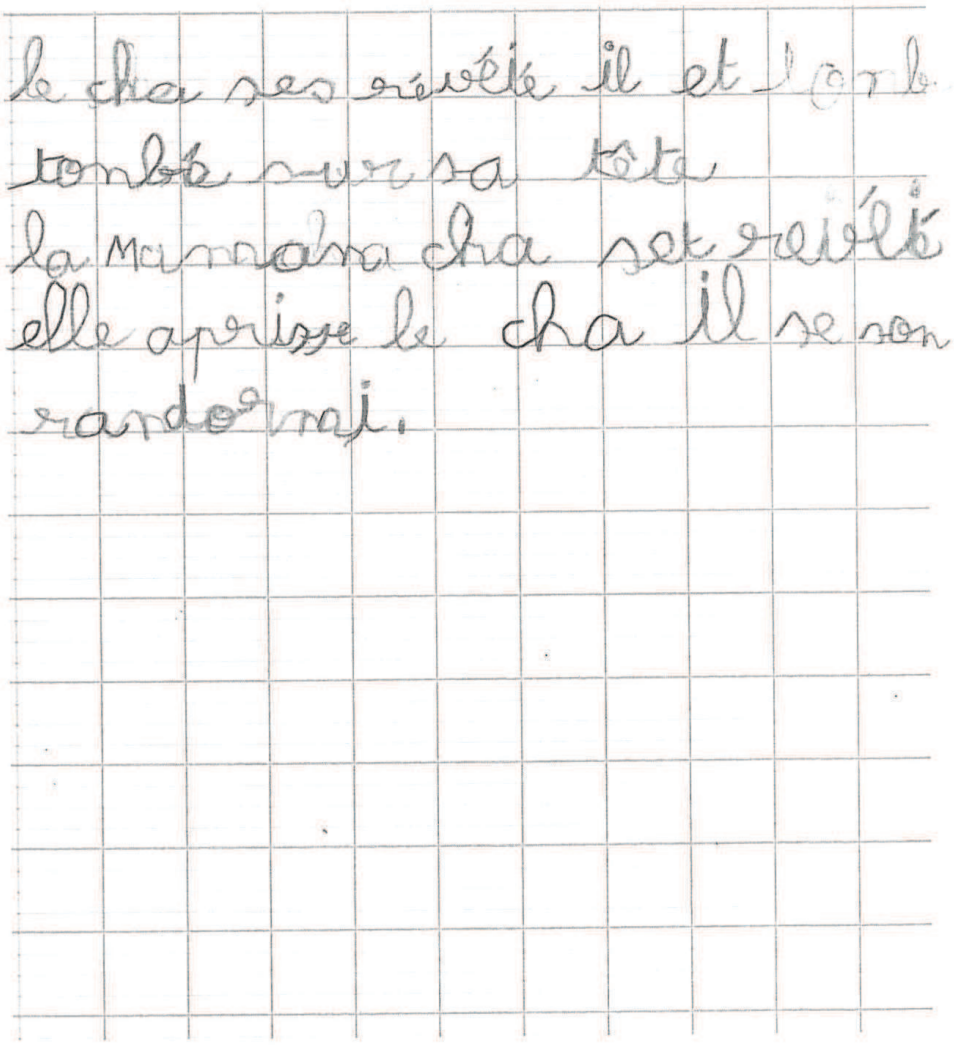

le chat s'en va le chat tombe le chat fé miaou
le chat va voir sa maman.

1596


le chat se rase il tond il
la maman chat se
le bébé chat est le
87
1666

il été 1 fois un chat et sa maman
dort même le chat se réveil et part
et boum le chat tombe sa maman se
réveil et va le chèrché elle è rassurer.

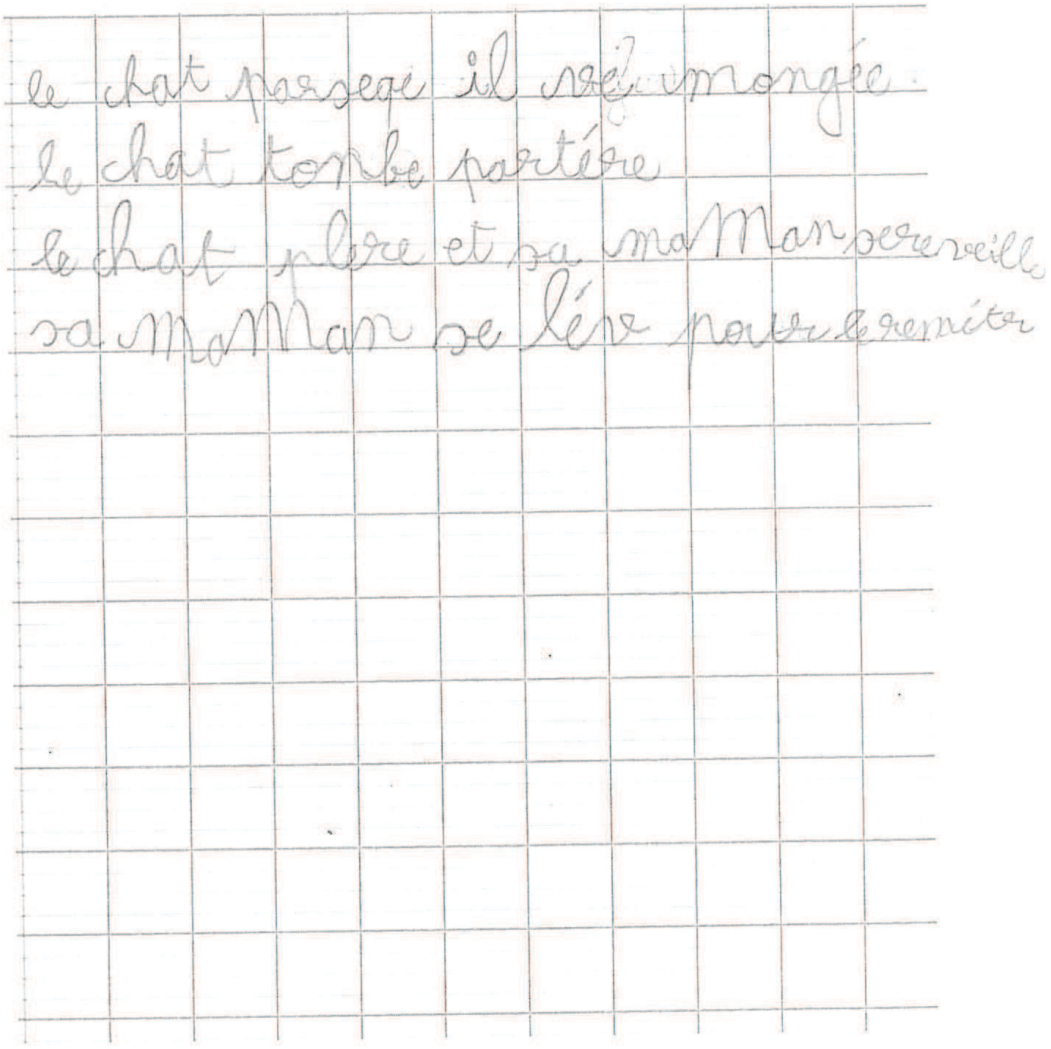
1956


le cha ses réveié il et l'orb
tombé sur sa tête
la Mamaha cha sek reveié
elle aprisse le cha il se son
randormi.

1986

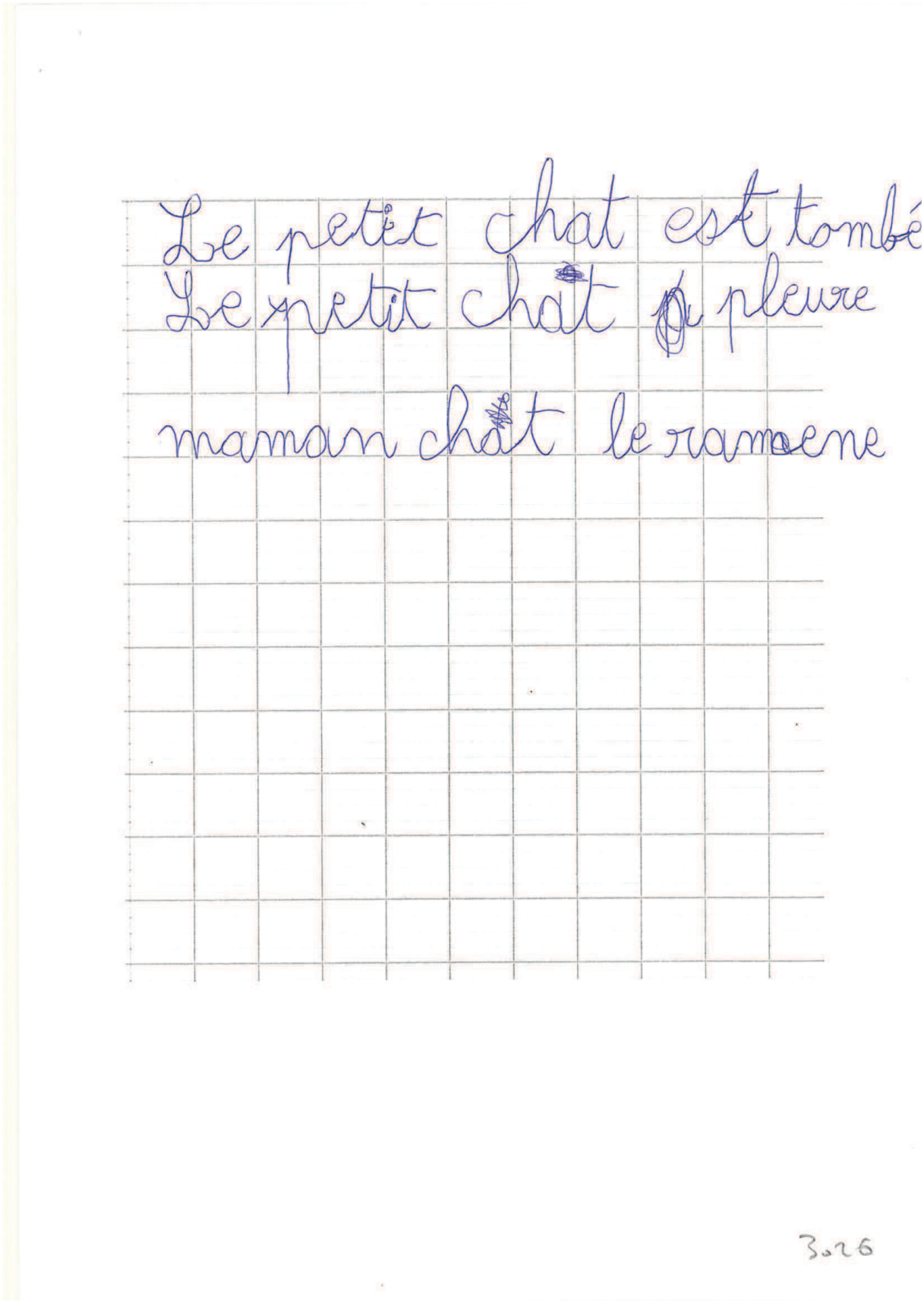
un petit chat se promène dans la nuit.
le petit chat se fémale.
et il pleure téfaire.
et il retrouve sa maman.

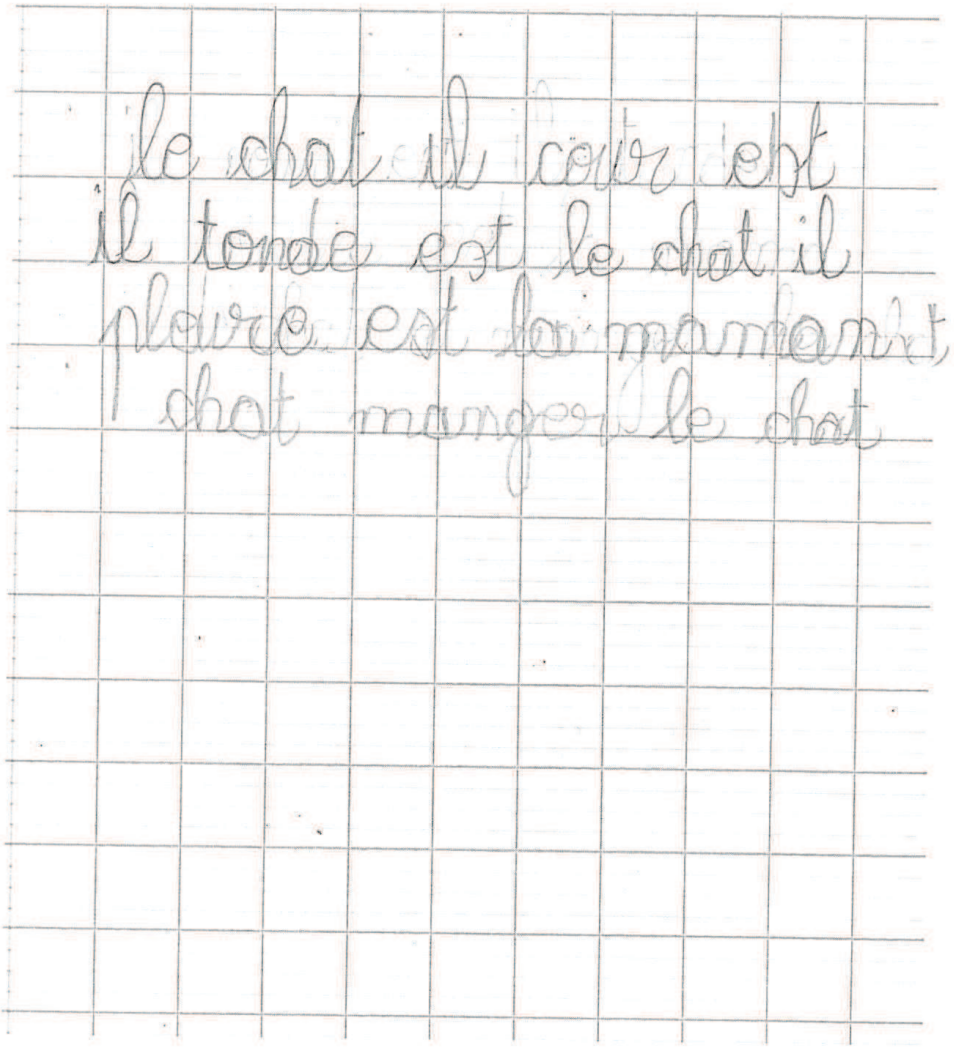
2976


le chat passage il ne mange
le chat tombe par terre
le chat pleure et sa maman se réveille
sa maman se lève pour le remettre

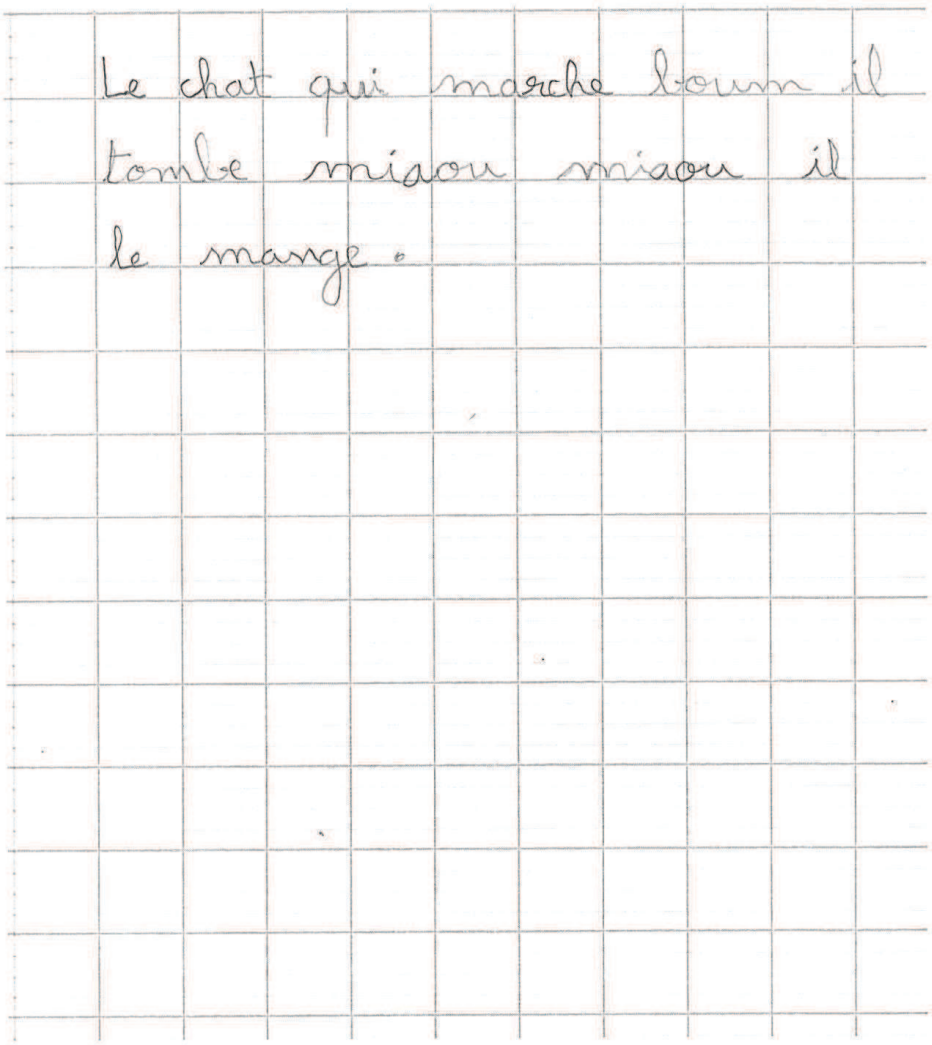
2986


le chat dar après le chat se can le chef
Miaou

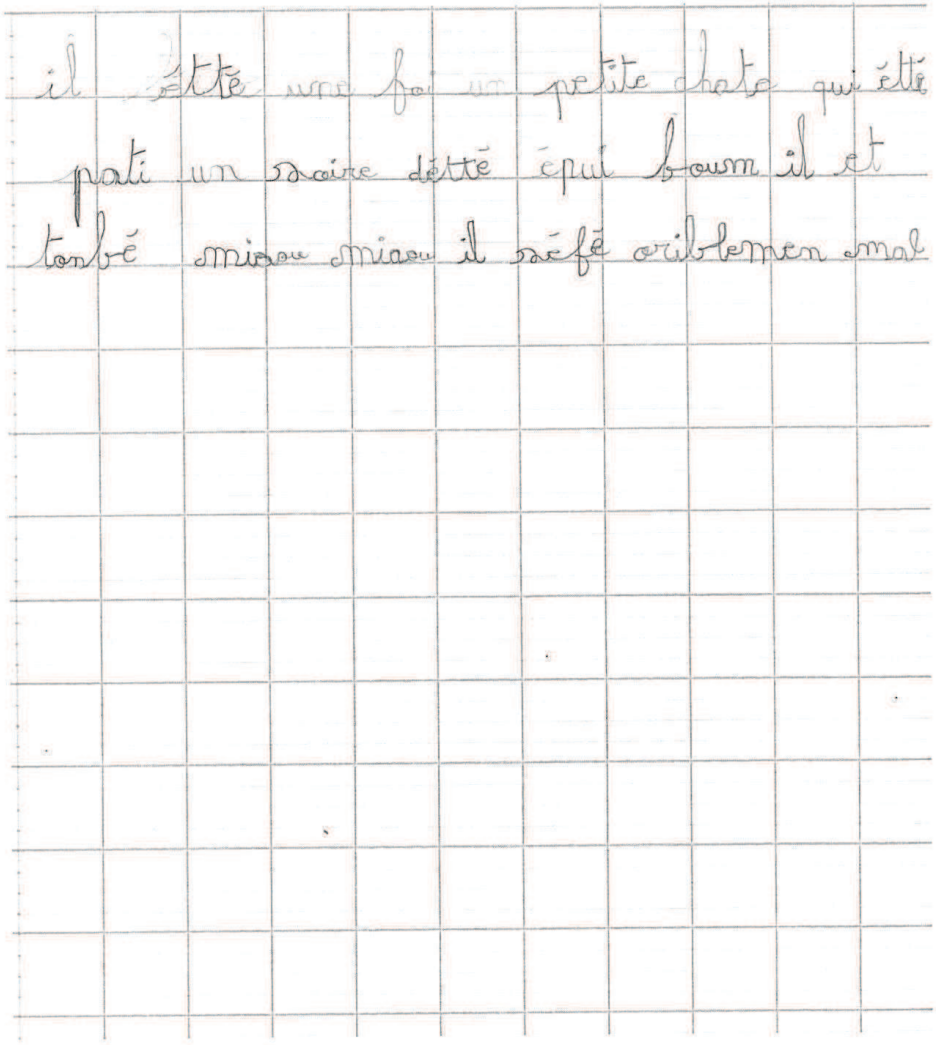

Le petit chat est tombé
Le petit chat pleure
maman chat le ramene
3.26


le chat il court
il tombe est le chat il
pleure est la maman
chat mange le chat

3.46



3.56


il étté une foi un petite chato qui étté
pati un soire déttè épui boum il et
tombé miaou miaou il séfé oriblemen mal

3.66

Transcription :

576 :
r r r r, bousm miao<letMF>u</letMF><revision/>
586 :
Le petit chat et partie de son lit. Mai boum il / tombe et il pleurer. Sa maman et sai fraire se / raivaya et il le vo<letMF>i</letMF>llér pleurer. Et miantenan / sai le matin. La mamans les porte pour les / sortir du lit. // Fin
1156 :
il eté diféran des sotre à prè / il tombe <letMF>é</letMF>t aprè sa fait / boum et aprè il fait miaou deux / foi et sa mère la trape.
1166 :
Il yavè un petie chat qui entandai le quri d'un / dinosore et il se <revision/>quogne la tête et il plere et il / va <revision/> chez sa maman.
1226 :
Le petite chat voulet ce promener mer / dincou elle er tomber elle mer sa maman / à en dus plerer mer sa maman a ramner / a coter elle.
1296 :
Aujourd'hui le petit chat se <revision/> / promène seul. pui dinquou / badabome. miaou ! miaou ! / pui la mamn se reveille. / et elle le pose sur plasse.
1336 :
les chat<revision/>on un chat et des sendu de la ma<revision/>man / <revision/>il et <letMF>t</letMF>ondé avec parse que le chaton. et il pleure<revision/> et il / <letMF>à</letMF>i révéille la maman. et il tonda de sa / maman et il plera. la maman chat pose les chaton.
1346 :
Le petit chat sanva pan dan cesa / maman dore. Le petit chat ne regade / par ou il va é il tonbe. Le petit chat / fai miaou miaou. La mamn va cheirei bèsbe.
1556 :
Le chat c'est éloinié de sa maman. / Le chat est tonbé. Le chat pleure. / La maman et venus le cerr
1596 :
le chat s'en va le chat tonb le chat fé miaou / le chat va vr sa maman.
1666 :
le chat se <revision/>reive il tonbil pler boucou / la maman chat se revil elle pran / le bé<letMF>b</letMF>é chat est le me sur une table
1956 :
il <revision/>étè<illisible/>tune fois un chat et sa maman / dorè m<letMF>a</letMF>i<m|n> le chat se révei et pare / et boum le chat tonba sa maman se / réveia et va le chèré elle è rasurer.
1986 :
le cha ses révéié il et <revision/> / tonbé sur sa tête // la maman<revision/> cha set reivéié / elle aprisse le cha il se son / randormi.
2976 :
un petit chat se <revision/> peméne dans la nui. / le petit chat se fé male. // et il pelere téfarre. // et il retruve sa maman.
2986 :
le chat parseqe il ve <revision/> mongée / le chat tonbe partére // le chat plere et sa maman sereveille / sa maman se lév pour le remétr
3006 :
le cha dar apér le cha se can le ch fe / Miaou
3026 :
Le petit chat est tombé / Le petit ch<revision/>at <revision/>a pleure / maman ch<revision/>at le ram<revision/>ene
3046 :
<revision/>le chat il cour est / <revision/>il tonde est le chat il / <revision/> pleure est la maman <revision/> / chat manger le chat
3056 :
Le chat qui marche boum il / tombe miaou miaou il / le mange.
3066 :
<revision/> il étté une foi un petite chate qui étté / pati un soire détté épui boum il et / tonbé miaou miaou il séfé oriblemen mal

Annexe 2 :

Consignes aux évaluateurs

Productions écrites – CP

(Extrait du livret de l'évaluateur de l'IFÉ)

B – Test collectif

3. Production d'écrit

Test de production d'une histoire sur la base de 4 images montrées par l'évaluateur.

En temps limité – 15 minutes (après les consignes).

a) Conditions de passation et matériel

Matériel évaluateur :
Chronomètre.
En annexe A, les 4 vignettes reproduites sur 4 feuilles au format A3 (feuilles séparées)
Aimants ou scotch pour afficher les vignettes au tableau.
Un exemplaire du livret élève pour montrer la page où ils écriront.
Matériel élève :
Un crayon papier et une gomme.
Dans le cahier, on trouve la page avec les lignes puis, au verso, les 4 vignettes reproduites.

Les élèves sont assis à leur place. L'enseignant est dans la classe ; il aide au bon déroulement de l'épreuve. Le prévenir de ne donner aucune aide aux élèves. On retire les affichages muraux susceptibles d'aider les élèves (le mot *chat*, des listes de verbes, etc.).

Les cahiers seront distribués une fois que l'évaluateur aura montré les 4 images.

b) Consigne

Quand les élèves sont tous attentifs, dire lentement : **« Aujourd'hui vous allez écrire chacun l'histoire d'un petit chat. Je vais vous montrer ce qui arrive à ce petit chat. Regardez bien comment ça commence. »**

Pour l'ordre des images, s'appuyer sur la série représentée dans le livret élève.

Montrer image 1 (format A3).

« Est-ce que vous le voyez, ce petit chat ? Oui, il est ici »

Le montrer du doigt. Puis bien montrer l'image à tous les élèves et l'afficher sur la partie la plus à gauche du tableau.

Si les élèves s'expriment, leur faire signe qu'ils ne doivent rien dire : **« Chut, regardez bien… On va regarder sans rien dire. Vous gardez toutes vos idées dans votre tête… »**

Montrer image 2 (format A3).

« Regardez bien la 2ème image… »

Si les élèves s'expriment, leur faire signe qu'ils ne doivent rien dire : **« Chut, regardez bien…»**

Bien montrer l'image à tous les élèves et l'afficher à droite de la première.

Montrer image 3 (format A3).

« Regardez bien la 3ème image… »

Si les élèves s'expriment, leur faire signe qu'ils ne doivent rien dire : **« Chut, regardez bien…»**

Bien montrer l'image à tous les élèves et l'afficher à droite de la deuxième image.

Montrer image 4 (format A3).

« Regardez bien la 4ème image… »

Si les élèves s'expriment, leur faire signe qu'ils ne doivent rien dire : **« Chut, regardez bien…»**

Bien montrer l'image à tous les élèves et l'afficher à droite des trois premières.

Les vignettes resteront au tableau pendant que les élèves écrivent.

« Vous avez bien regardé, vous avez bien dans votre tête l'histoire de ce petit chat ? »

Il est important de leur laisser le temps de « mettre l'histoire dans leur tête ».

Montrer ensuite la feuille du cahier sur laquelle ils vont écrire (lignes Seyes).

« Vous allez écrire cette histoire ici. Si vous avez oublié l'histoire, vous pouvez retourner la feuille pour retrouver les dessins. »

« Vous avez 15 minutes pour ce travail. Vous allez travailler seul ; personne ne vous aidera, par exemple à écrire un mot.

Distribuer les cahiers, bien faire repérer la page d'écriture et déclencher le chronomètre quand tous les élèves sont prêts. Avertir les élèves lorsqu'il restera 5 minutes. Ramasser les cahiers au bout de 15 minutes.

Durant le déroulement de cette épreuve, faire les réponses les plus neutres possible, ne pas donner d'aide aux élèves.

Quand les élèves ont fini, leur dire de se reposer, par exemple en posant la tête sur les avant-bras.

- Ne pas autoriser le coloriage des vignettes qui pourrait inciter certains élèves à interrompre leur tâche d'écriture.

- Ne pas non plus autoriser les élèves à se lever et aller lire un livre.

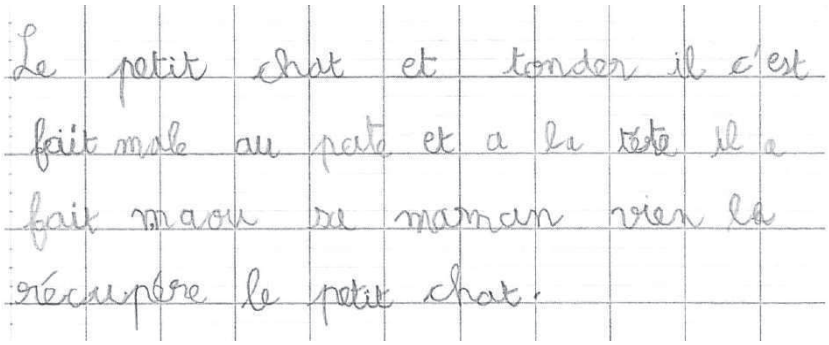
Annexe 3 :

Encodage et problèmes de numérisation

Productions écrites – CP

Marqueur de retour à la ligne :

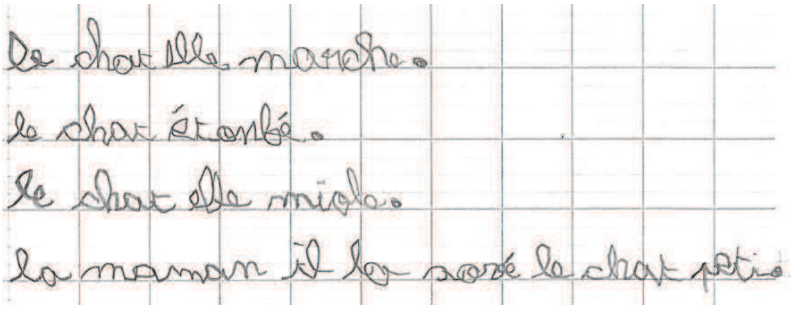
/ : Si le retour à la ligne est dû à la fine ligne sur la page



(Production **1354**)

L'exemple ci-dessus sera transcrit :
Le petit chat est tonder il c'est / fait male au pate et a la téte il a / fait maou sa maman vien le / récupére le petit chat.

// : Si le retour à la ligne est jugé volontaire et non contraint par le support

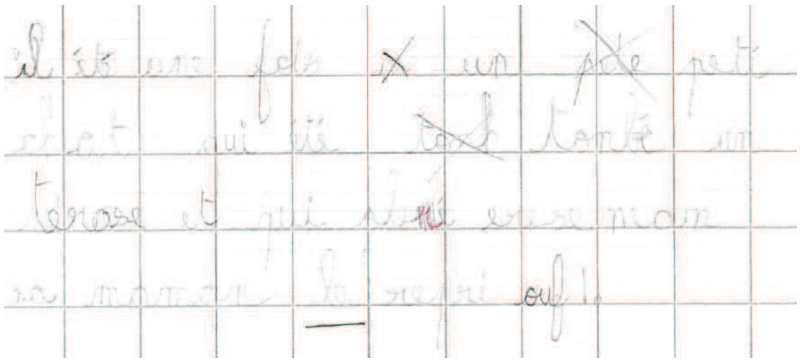


(Production **1361**)

L'exemple ci-dessus sera transcrit :
le chat elle marche. // le chat étonbé. // le chat elle miole. // la maman il la sové le chat peti.

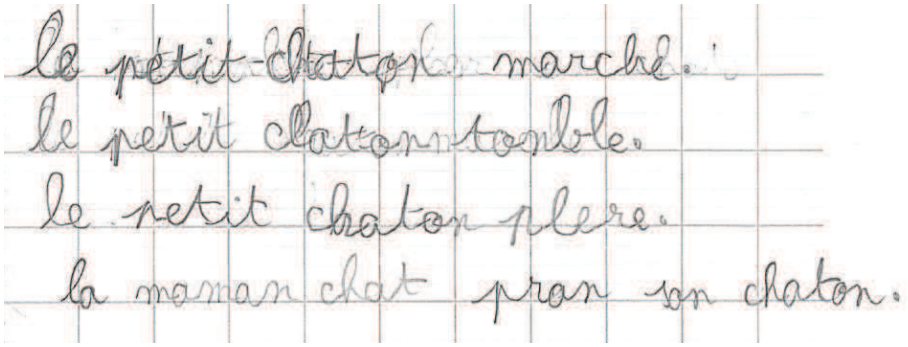
Marques de révision :

<revision/> : Pour tout élément comportant une rature, des traces de gomme ou des traces de réécriture.



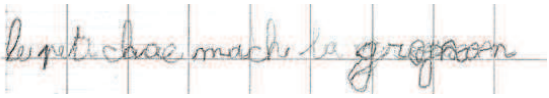
(Production **1560**)

L'exemple ci-dessus sera transcrit :
il été une fois <revision/> un <revision/> peti / chat qui été <revision/> tonbé une / terase et qui ple<revision/>ré ereseman / sa maman le repri ouf !.



(Production **1343**)

L'exemple ci-dessus sera transcrit :
<revision/> le petit chaton marche. // <revision/> le petit chaton tonble. // le petit chaton plere. // la maman chat pran son chaton.



(Production **2533**)

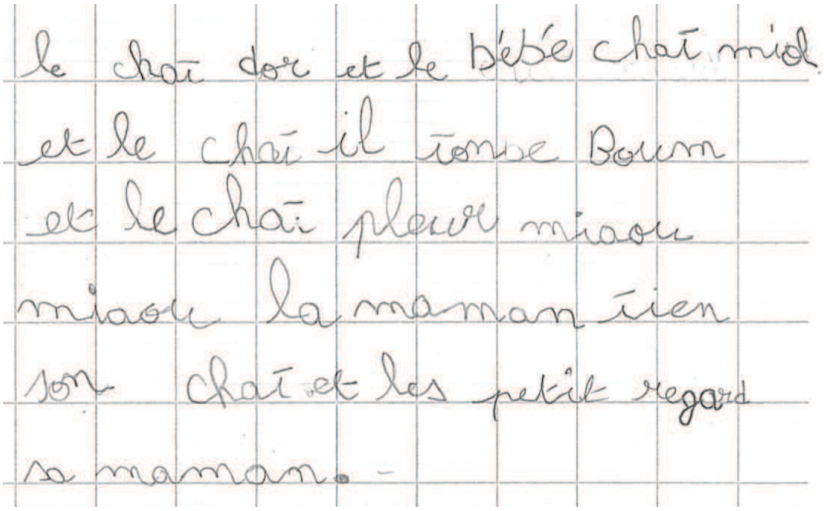
L'exemple ci-dessus sera transcrit :
le peti chae mach la <revision/><illisible/

Soulignement et les lettres entourées :

Le soulignement, tout comme les lettres ou mots entourés, étant parfois le fait du relecteur il n'est pas toujours possible de faire la différence entre un soulignement produit par l'élève de celui produit par l'adulte, nous n'en tiendront pas compte.

Lettres mal formées et segments illisibles :

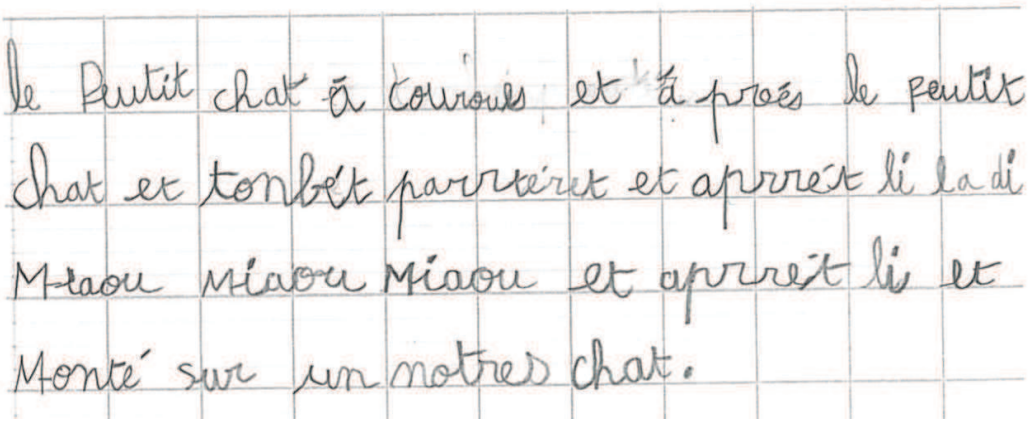
<letMF>x</letMF> : Pour les lettres que l'on peut déduire du contexte mais qui sont mal formées



(Production **1297**)

Cette production sera transcrite :
le cha<letMF>t</letMF> dor et le bébé cha<letMF>t</letMF> miol / et le cha<letMF>t</letMF> il <letMF>t</letMF>onbe Boum / et le cha<letMF>t</letMF> pleur miaou / miaou la maman <letMF>t</letMF>ien / son cha<letMF>t</letMF> et les petit regard / sa maman

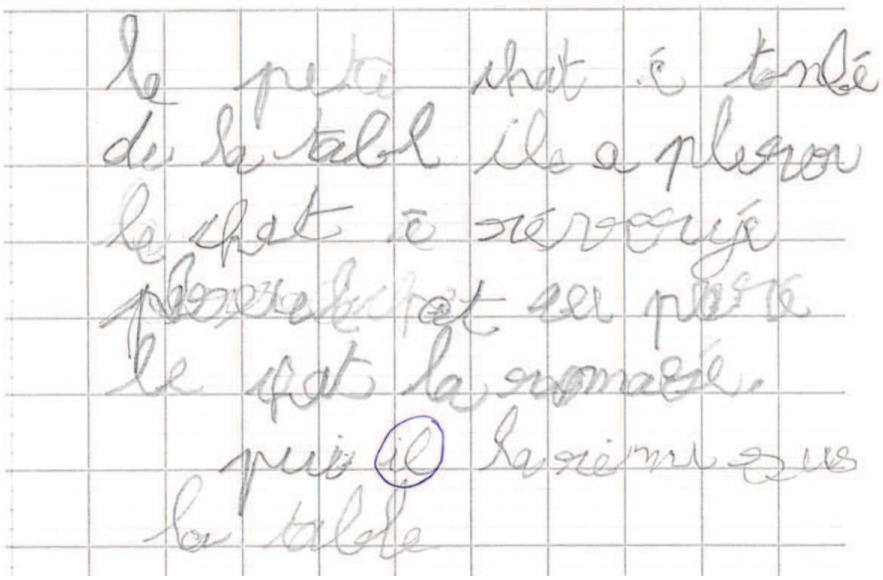
Ce marqueur est également utilisé pour indiquer un choix d'accent fautif et qui n'existe pas dans la langue à l'exemple de la production suivante :



(Production **1230**)

Dans cette production, les a accentués sont transcris avec un accent aigu et non grave, ce caractère n'existe pas en français. La transcription sera la suivante :
le Peutit chat <letMF>à</letMF> couroues<revision/> et <revision/> <letMF>à</letMF> preés le peutit / chat et tonbét parrtéret et aprrét li la di / Miaou Miaou Miaou et aprrét li et / Monté sur un notres chat.

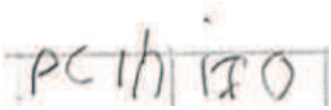
<illisible/ > : Pour les segments illisibles



(Production **1341**)

Cette production sera transcrite :
le pe<letMF>t</letMF>i chat é tonbé / de la tabl ile a plerer / le chat <letMF>ê</letMF> réveryé / <illisible/> <revision/> at <illisible/> pl<letMF>e</letMF>ré / le chat la ramasé. / pui il la remi sus / la table

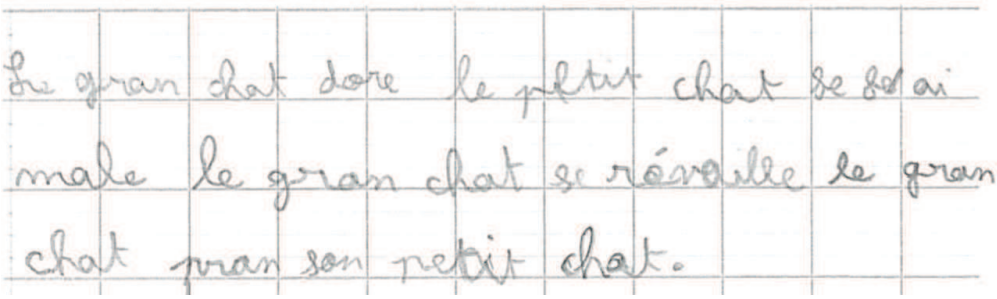
<letDet/> : Pour les suites de lettres détachées dont on arrive à extraire ni mots ni sens



(Production **1553**)

Pour cette production, seule la balise <letDet/> sera mentionnée.

<x|y> : Souvent le contexte permet de discriminer une lettre peu lisible. Lorsque ce n'est pas le cas, une balise permettra de donner les différentes possibilités envisagées.



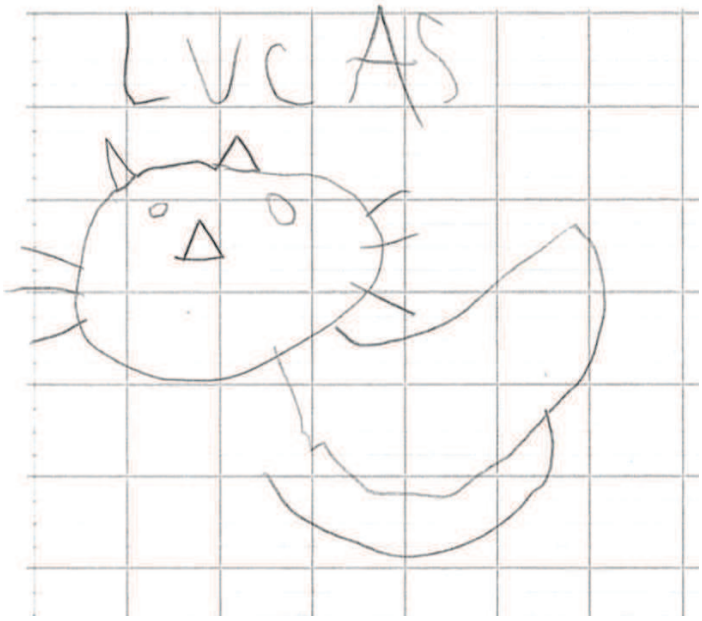
(Production **1340**)

Dans cette production, le doute persiste entre la lettre *e* et la lettre a pour « réveille », la transcription sera donc :
Le gran chat dore le p<letMF>e</letMF>tit chat se fe ai / male le gran chat se rév<a|e>ille le gran / chat pran son <revision/>petit chat.

Dessin et prénom

<dessin/> : présence de dessin dans la production

<prenom/> : écriture du prénom de l'enfant par l'enfant lui-même en place de la production



(Production **1288**)

Pour cette production, il sera fait mention des balises : <prenom/> // <dessin/>.

Annexe 4 :

Formes et lemmes présents dans un sous-corpus corrigé de 40 productions

Formes	Nombre d'occurrences
chat	74
il	67
le	65
et	55
maman	44
sa	31
se	26
petit	23
miaou	22
est	20
Le	19
la	19
un	18
tombe	16
pleure	13
a	12
fait	12
de	10
bébé	10
boum	10
va	10
réveille	10
s'	9
sur	9
elle	8
mais	8
était	8
une	7
les	6
fois	6
chercher	6
tombé	6
après	6
coup	6
ses	6
mal	5
promène	5
chaton	5
d'	5
l'	5
tomba	5
que	5
dort	5
en	5
petits	5
avec	4
miaule	4
qui	4
pleura	4
puis	4
par	3
ramène	3
tête	3
pleurer	3
Maman	3
parce	3
tout	3
bébés	3
réveilla	3
chats	3
mange	3
chatte	3
La	3
pour	3
boire	3
regarde	2
porte	2
pleuré	2
pendant	2
voulait	2
c'	2
son	2
chemin	2
table	2
ça	2
terre	2
prend	2
rrrr	2
des	2
Une	2
pas	2
lit	2
dans	2
trois	2
y	2
marche	2
réveillé	2
parti	2
du	2
pose	2
mère	2
éloigne	2
promener	2
part	2
court	2
chatons	2
dormaient	2
trébucher	2
dit	2
seul	2
attrape	2
sont	2
tombée	1
cri	1
côté	1
srolé	1
vol	1
allant	1
quand	1
rebord	1
avait	1
très	1
où	1
copains	1
voyaient	1
matin	1
beaucoup	1
remettre	1
soir	1
sorti	1
réveillèrent	1

ne	1
t'	1
nuit	1
profite	1
aider	1
autres	1
quitte	1
descendu	1
ensuite	1
jour	1
rendormi	1
sortir	1
balade	1
soudain	1
trottoir	1
deux	1
cogne	1
ramasse	1
horriblement	1
ramener	1
maintenant	1
bouche	1
entendu	1
cinq	1
partit	1
retrouve	1
rappelle	1
vers	1

veut	1
aller	1
assis	1
mangé	1
pris	1
venue	1
ti	1
garde	1
tomber	1
nid	1
Aujourd'hui	1
ils	1
Fin	1
frères	1
place	1
leur	1
lève	1
vu	1
vient	1
gros	1
badaboum	1
toit	1
récupère	1
rassurée	1
entend	1
rmr	1
mery	1
eau	1

Un	1
entendait	1
sauf	1
réveiller	1
Ronron	1
manger	1
arrête	1
marcher	1
différent	1
chez	1
lui	1
fort	1
r	1
content	1
peu	1
petite	1
dinosaure	1
été	1
enfant	1
enfants	1
éloigné	1
regardent	1
marchait	1
met	1
alla	1

Lemmes	Nombre d'occurrences
le	114
chat	77
il	68
et	55
maman	47
son	39
se	35
être	30
tomber	29
petit	29
un	28
pleurer	22
miaou	22
réveiller	17
de	15
avoir	13
aller	13
faire	12
bébé	12
boum	10
sur	9
mais	8
elle	8
promener	7
dormir	7
chaton	7
coup	6
chercher	6
après	6
que	5
manger	5
mal	5
en	5
ramener	4
qui	4
puis	4
partir	4
miauler	4
du	4
avec	4
à	4
vouloir	3
tout	3
tête	3
regarder	3
prendre	3
pour	3
parce	3
par	3
marcher	3
la|le	3
fois	3
foi|fois	3
entendre	3
éloigner	3
boire	3
y	2
voir	2
venir	2
trois	2
trébucher	2
terre	2
table	2
sortir	2
seul	2
rrrr	2
poser	2
porter	2
pendant	2
pas	2
mère	2
lit	2
enfant	2
dire	2
dans	2
courir	2
chemin	2
chatte	2
cela	2
ce	2
vol	1
vers	1
trottoir	1
très	1
toit	1
te	1
soudain	1
soir	1
sauf	1
Ronron	1
retrouver	1
rendormir	1
remettre	1
récupérer	1
rebord	1
rassurer	1
rappeler	1
ramasser	1
quitter	1
quand	1
profiter	1
place	1
peu	1
part	1
où	1
nuit	1
nid	1
ne	1
mettre	1
matin	1
marche	1
maintenant	1
lui	1
lever	1
leur	1
jour	1
horriblement	1
gros	1
garder	1
frère	1
fort	1
fin	1
été	1
ensuite	1
ébé	1
eau	1
dinosaure	1
différent	1
deux	1
descendre	1
cri	1
côté	1
copain	1
content	1
cogner	1
cinq	1
chez	1
chat|chatte	1
bouche	1
beaucoup	1
balader	1
badaboum	1
autre	1
aujourd'hui	1
attraper	1
attrape	1
asseoir	1
arrêter	1
aider	1
<unknown>	5

Annexe 5 :

Observations à partir d'un sous-corpus de 17 productions

Omission

Lettres muettes

Lettre ou groupe de lettres	N° production	Forme erronée	Forme attendue
t final de conjugaison			
t	1156	eté	était
t	1956	été	était
t	1166	yavè	y avait
t	1166	entandai	entendait
t	1346	fai	fait
t	1596	fé	fait
t	2976	fé	fait
t	3006	fe	fait
t	3066	séfé	s'est fait
t	3066	étté	était
t	3066	étté	était
t	1986	son	sont
t	2986	ve	veut
t	3006	dar	part
t	3046	cour	court
t final invariable			
t	1156	diféran	différent
t	1346	pan dan	pendant
t	1986	cha	chat
t	1986	cha	chat
t	1986	cha	chat
t	3006	cha	chat
t	3006	cha	chat
t	3006	ch	chat
t	2976	nui	nuit
t	2976	téfarre	très fort
t	3036	peti	petit
t	3036	peti	petit
t	3036	petichat	petits chats
t	3066	oriblemen	horriblement
s final de pluriel			
s	1156	sotre	autres
s	1336	chaton	chatons
s	3036	petichat	petits chats
s	1986	il	ils
s	1986	randormi	rendormi
s final invariable			
s	1156	à prè	après
s	1156	aprè	après
s	1156	aprè	après
s	3006	apèr	après
s	1156	foi	fois
s	3066	foi	fois
s	1226	mer	mais
s	1226	mer	mais
s	1226	mer	mais
s	1226	mé	mais
s	1596	vr	vers
s	2976	téfarre	très fort
s	3066	épui	et puis
p final invariable			
p	1226	dincou	d'un coup
p	1666	boucou	beaucoup
e final de féminin sur participe passé			
e	1226	tomber	tombée
e	1956	rassurer	rassurée
e final de conjugaison			
e	1596	tonb	tombe
e	1666	tonbil	tombe il
e	1166	pler	pleure
e	1666	revil	réveille
e	1956	révei	réveille
d final de conjugaison			
d	1666	pran	prend
conjugaison 3e du pluriel			
ent	1956	dorè	dormaient
Semi-voyelle			
ll	1956	révei	réveille
ll	1956	réveia	réveilla
ll	1986	révéié	réveillé
milieu de mot			
e	1226	ramner	ramené
u	2986	parseqe	parce que
h	3066	oriblemen	horriblement

Lettres prononcées

Lettre ou groupe de lettres	N° production	Forme erronée	Forme attendue
ch	1346	cheirei	chercher
ch	1556	cerr	chercher
ch	1956	chèré	chercher
ll	1666	reive	réveille
ll	1986	reivé	réveillé
m	1956	dorè	dormaient
r	1346	regade	regarde
r	2976	téfarre	très fort

r	3066	pati	parti
a	3006	ch	chat
e	1596	vr	vers
é	3026	tomb	tombé
ei	1666	reive	réveille
ei	1986	reivé	réveillé

Dans diphtongue

a	1346	mamn	maman
e	1556	cerr	chercher
e	1666	revil	réveille
o	2976	retruve	retrouve
n	1986	mama	maman
h	1556	cerr	chercher

Doublement consonne

Lettre ou groupe de lettres	N° production	Forme erronée	Forme attendue
f	1156	diféran	différent
p	1156	la trape	l'attrape
l	1666	revil	réveille
r	2986	partére	par terre
r	3066	oriblemen	horriblement
t	2986	remétr	remettre

Syllabe

Lettre ou groupe de lettres	N° production	Forme erronée	Forme attendue
ten	1226	en dus	entendu

Accent

Lettre ou groupe de lettres	N° production	Forme erronée	Forme attendue
é	1156	eté	était
é	1336	révéille	réveillé
é	1666	revil	réveille
é	2986	sereveille	se réveille
é	3036	tanbe	tombé
è	3026	rama ene	ramène
à	1226	a	à
ô	1226	coter	côté
ù	1346	ou	où

Insertion

Lettres muettes

Lettre ou groupe de lettres	N° production	Forme erronée	Forme attendue
s	1226	en dus	entendu
e	2976	male	mal
e	2976	pelere	pleure
e	2976	téfarre	très fort
e	2986	mongée	monger
e	3066	soire	soir

Lettres prononcées

Lettre ou groupe de lettres	N° production	Forme erronée	Forme attendue
q	1336	parseq	parce
i	1336	ài	après
s	1986	aprisse	a pris
s	3006	sse	se
a	3026	rama ene	ramène
muette ?			
e	3066	petite	petit
e	3066	chate	Chat
e	1986	aprisse	a pris
r	3046	manger	mange

Doublement consonne

Lettre ou groupe de lettres	N° production	Forme erronée	Forme attendue
r	2976	téfarre	très fort
t	3066	étté	était
t	3066	étté	était
t	3066	détté	d'été
e	1986	aprisse	a pris

Accent

Lettre ou groupe de lettres	N° production	Forme erronée	Forme attendue
a -> à	1156	à prè	après
a -> à	1336	ài	après
a -> ä	1956	mäim	mais
e -> é	1156	ét	et
e -> é	1986	révéié	réveillé
e -> é	2986	partére	par terre
e -> é	2986	remétr	remettre

Liaison

Lettre ou groupe de lettres	N° production	Forme erronée	Forme attendue	contexte gauche
s	1156	sotre	autres	des

Substitution

Phonologie respectée

Lettre ou groupe de lettres	N° production	Forme erronée	Forme attendue
Terminaison en ai + lettre muette			
Verbe ai + t			
ai -> é [éè]	1156	eté	était
ai -> è [éè]	1956	étè	était
ai -> è [éè]	1166	yavè	y avait
ai -> et [éè]	1226	voulet	voulait
ai -> é [éè]	1596	fé	fait
ai -> é [éè]	2976	fé	fait
ai -> é [éè]	1226	mé	mais
ai -> é [éè]	3066	étté	était
ai -> é [éè]	3066	étté	était
ai -> é [éè]	3066	séfé	s'est fait
Conjonction ai +s			
ai -> er [éè]	1226	mer	mais
ai -> er [éè]	1226	mer	mais
ai -> er [éè]	1226	mer	mais
Forme verbale est			
est -> er [e]	1226	er	est
est -> et [e]	1336	et	est
est -> et [e]	1556	et	est
est -> et [e]	1986	et	est
est -> et [e]	3046	et	est
est -> et [e]	3046	et	est
est -> et [e]	3046	et	est
est -> et [e]	3066	et	est
est -> è [e]	1956	è	est
est -> es [e]	1986	ses	s'est
est -> et [e]	1986	set	s'est
est -> é [e]	3066	séfé	s'est fait
Conjonction et			
et -> est [e]	1556	est	et
et -> é [e]	3066	épui	et puis
et -> é [e]	1346	é	et
Participes passés en é			
é -> er [e]	1226	tomber	tombée
er -> é [e]	1226	ramner	ramené
é -> er [e]	1956	rassurer	rassurée
Infinitifs en er			
er -> é [e]	2986	mongée	manger
er -> ei [e]	1346	cheirei	chercher
er -> é [e]	1956	chèré	chercher
Finales de nom en é			
er -> é [e]	1226	coter	côté
é -> e [e]	1346	beèbe	bébé
é -> ei [e]	1666	reive	réveille
é en milieu de nom			
é -> ei [e]	1986	reivé	réveillé
é -> eè [e]	1346	beèbe	bébé
Son [ɜ] écrit e			
e -> ei [ɜ]	1346	cheirei	chercher
e -> è [ɜ]	1956	chèré	chercher
Son [œ] écrit eu			
eu -> e [œ]	1166	plere	pleure
eu -> e [œ]	2986	plere	pleure
eu -> e [œ]	1226	plerer	pleurer
eu -> e [œ]	1336	plera	pleura
eu -> e [œ]	1166	pler	pleure
eu -> e [œ]	2976	pelere	pleure
eu -> e [œ]	2986	ve	veut
Son [ɔ]			
au -> o [ɔ]	1166	dinosore	dinosaure
Son [ã] écrit en			
en -> an [ã]	1156	diféran	différent
en -> an [ã]	1166	entandai	entendait
en -> an [ã]	1346	sanva	s'en va
en -> an [ã]	1346	pan dan	pendant
en -> an [ã]	1666	pran	prend
en -> an [ã]	1986	randormi	rendormi
Son [ɜ̃] écrit un			
un -> in [ɜ̃]	1226	dincou	d'un coup
Son [ɔ̃] écrit om			
om -> on [ɔ̃]	1336	tondé	tombé
om -> on [ɔ̃]	1336	tonda	tomba
om -> on [ɔ̃]	1346	tonbe	tombe
om -> on [ɔ̃]	1556	tonbé	tombé
om -> on [ɔ̃]	3066	tonbé	tombé
om -> on [ɔ̃]	1596	tonb	tombe
om -> on [ɔ̃]	1666	tonbil	tombe il
om -> on [ɔ̃]	1956	tonba	tombe
om -> on [ɔ̃]	1986	tonbé	tombé
om -> on [ɔ̃]	2986	tonbe	tombe
om -> on [ɔ̃]	3046	tonde	tombe
Consonnes			
c -> s [s]	1336	parseq	parce
c -> s [s]	2986	parseqe	parce
s -> c [s]	1346	cen	sa
s -> c [s]	1556	c'	s'
ç -> s [s]	1156	sa	ça
sc -> ss [s]	1336	des sendu	descendu
c -> qu [k]	1166	quri	cri
c -> qu [k]	1166	quogne	cogne
gn -> ni	1556	éloinié	éloigné

Phonologie proche

Lettre ou groupe de lettres	N° production	Forme erronée	Forme attendue
b -> d	1336	tondé	tombé
b -> d	1336	tonda	tomba
b -> d	3046	tonde	tombe
eau -> ou	1666	boucou	beaucoup
ai -> e	3006	fe	fait
a -> en [a]	1346	cen	sa

Doublement consonne

Lettre ou groupe de lettres	N° production	Forme erronée	Forme attendue
f	1156	diféran	différent
p	1156	la trape	l'attrape
l	1666	revil	réveille
r	2986	partére	par terre
r	3066	oriblemen	horriblement
t	2986	remétr	remettre

Graphie proche

Lettre ou groupe de lettres	N° production	Forme erronée	Forme attendue
an -> on	2986	mongée	manger
om -> an	3036	tanbe	tombé
o -> a	2976	téfarre	très fort

Accent

Lettre ou groupe de lettres	N° production	Forme erronée	Forme attendue
è -> é	2976	proméne	promène
è -> é	2976	téfarre	très fort
è -> é	2986	léve	lève

Autre

Lettre ou groupe de lettres	N° production	Forme erronée	Forme attendue
s -> m	1956	mäim	mais
p -> d	3006	dar	part

Lettres muettes

Lettre ou groupe de lettres	N° production	Forme erronée	Forme attendue
t -> e	1166	petie	petit
t -> e	1346	dore	dort
t -> e	1956	pare	part
s -> r	1346	par	pas
e ->s	1556	venus	venue

Inversion

Inversion

N° production	Forme erronée	Forme attendue
3006	apèr	après

Segmentation

Sous segmentation

N° production	Forme erronée	Forme attendue
Collocation et expressions figées		
1166	yavè	y avait
1346	sanva	s'en va
2986	partére	par terre
3066	épui	et puis
1226	dincou	(tout) d'un coup
Lexème unique		
2986	parseqe	parce que
Autre		
1666	tonbil	tombe il
1986	aprisse	a pris
2976	téfarre	très fort
2986	sereveille	se réveille
3036	petichat	petits chats
3066	séfé	s'est fait

Elision

Lettre ou groupe de lettres	N° production	Forme erronée	Forme attendue
l'	1156	la trape	l'attrape
d'	1226	dincou	d'un coup
d'	3066	détté	d'été
s'	1346	sanva	s'en va
s'	1986	ses	s'est
s'	1986	set	s'est
s'	3066	séfé	s'est fait

Sur segmentation

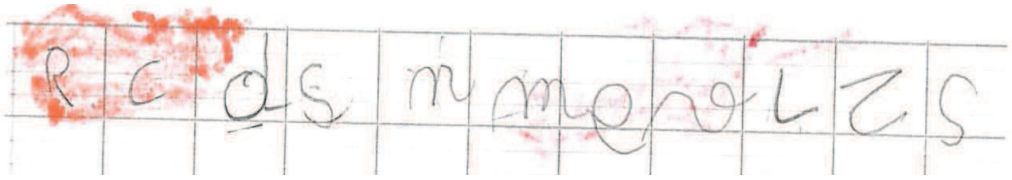
N° production	Forme erronée	Forme attendue
Forme identifiée		
1156	à prè	après
1226	en dus	entendu
1336	des sendu	descendu
1156	la trape	l'attrape
Forme inexistante		
1346	pan dan	pendant
3026	rama ene	ramène

Annexe 6 :

Productions pré-syllabiques, syllabiques et syllabo-alphabétiques

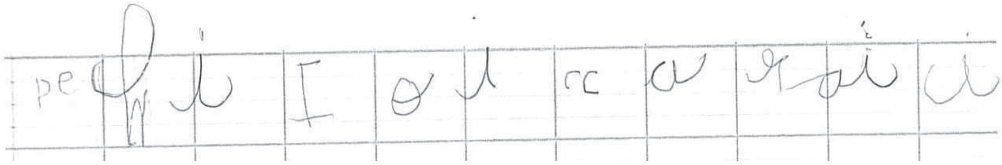
Productions pré-syllabiques et syllabiques, annotées <letDET/> :

1170 :



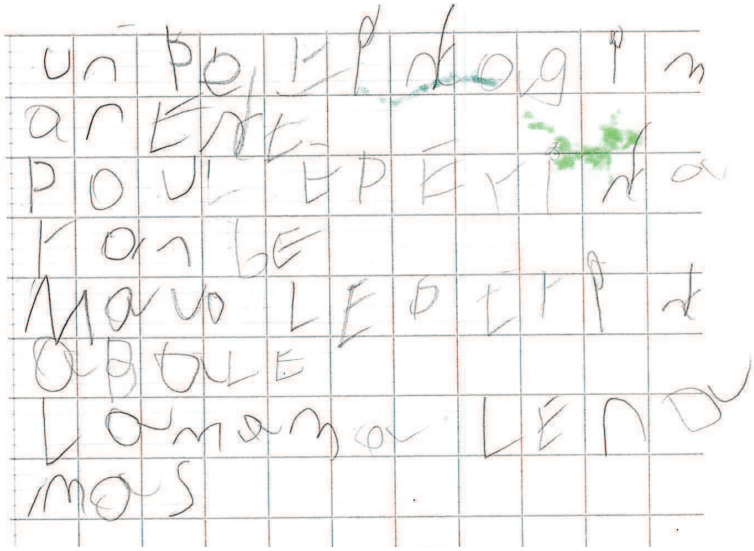
1553 :

1671 :

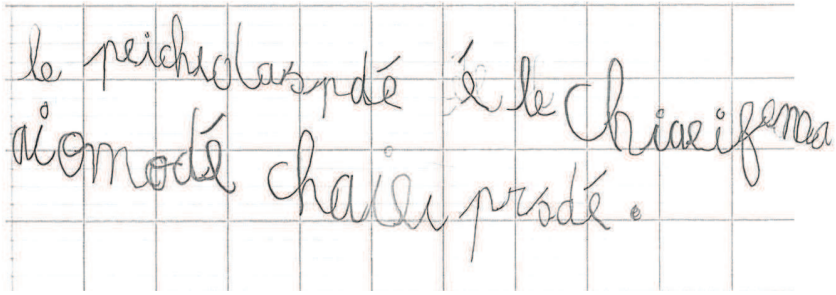


Exemples de productions syllabo-alphabétiques :

2367 :



1983 :



Annexe 7 :
Tableau des correspondances phonographiques, Catach, 1978

Archigraphèmes		Graphèmes de base	Pourcentage d'utilisation
		Voyelles	
A		a	92
E	[e]	e + é	99
	[ɛ]	(e) + è	68
		ai	30
I		i	99
O		o	75
		au	21
		eau	3
U		u	100
EU		eu	93
		(e)	
OU		ou	98
AN		an	44
		en	47
IN		in	45
		(en)	47
ON		on	92,8
UN		un	97
		Semi-voyelles	
OI		oi	100
OIN		oin	100
IL(L),Y		(i)	86
		l	
		il(l)	10
		y	3
		Consonnes	
P		p	100
B		b	100
T		t	99
D		d	100
C		c + qu	98
G		g + gu	100
F		f	95
V		v	100
S		s + ss	69
		(c) + ç	26
Z		(s intervocalique)	90
		z	10
X		x	84
CH		ch	100
J		j	49
		(g) + ge	51
L		l	100
R		r	100
M		m	100
N		n	100
GN		gn	100

Annexe 8 :

Réflexions et choix des phénomènes à réécrire dans une proposition de réécriture normée

Productions écrites – CP

1) **Notations spécifiques**

<nonfini> : Mots ou phrases non achevés en fin de production

et di maou maou qui **r** (**1228**, *et dit « miaou miaou » qui r**<nonfini>***)

<incomprehensible> : Segments qui échappent à la compréhension

l ti chat **pas c ese** bébe cha (**1363**, *le petit chat **<incomprehensible>** bébé chat*)

2) **Eléments réécrits**

Segmentation en mots :

il l**ep**ren (**563**, *il l**e p**rend*)

Orthographe lexicale :

Il **e**tait une foi**e** (**568**, *Il **é**tait une foi**s***)

Phénomènes d'accords :

et les chaton aussi (**1143**, *et les chaton**s** aussi*)

Répétitions :

le petit chat **le patit chat** fé dodo (**1598**, *le petit chat fait dodo*)

Rétablissement des marques de l'écrit :

- Ponctuation, majuscules et guillemets

soudin il tonbe **»** il dit **»** miaou **«** (**587**, ***S**oudain**,** il tombe**.** **I**l dit **«** miaou **».***)

- Négation

mai il a pas vus le trautoir (**562**, *Mais il **n'**a pas vu le trottoir.*)

Rétablissement de certaines structures syntaxiques :

- Pronom relatif

il ettai tune fou un peti chat et tai sor. ti (**1228**, *Il était une fois un petit chat **qui** était sorti*)

3) **Ambigüité ou possibilités diverses**

En cas d'ambigüité, la proposition la plus probable est choisie.

il marche **sur au bor** de la plake. (**1352**, *Il marche **sur le bord** de la plaque.*)

4) **Eléments non réécrits**

Temps des verbes :

il tomb**e** et il pleur**er**. (**586**, *Il tomb**e** et il pleu**rait**.*)

Attestation multiple des groupes nominaux :

il denre **les chat**. (**1290**, ***Ils*** *dorment* ***les chats***.)

Utilisation « abusive » des connecteurs *et* :

le chat quoure **é** il tonbe **é** il plére **é** sa maman lui fai un qualin **é** le chat na plumale (**573**, *Le chat court* ***et*** *il tombe* ***et*** *il pleure* ***et*** *sa maman lui fait un câlin* ***et*** *le chat n'a plus mal.*)

Annexe 9 :

Correspondances des transcriptions de phonèmes entre notation API et format LIA

Format API	Format LIA
/p/	pp
/b/	bb
/t/	tt
/d/	dd
/k/	kk
/g/	gg
/f/	ff
/v/	vv
/s/	ss
/z/	zz
/ʃ/	ch
/ʒ/	jj
/l/	ll
/r/	rr
/m/	mm
/n/	nn
/a/	aa
/e/	ei
/ɛ/	ai
/i/	ii
/o/	au
/ɔ/	oo
/y/	uu
/ø/	eu
/œ/	oe
/ə/	ee
/u/	ou
/ɑ̃/	an
/ɛ̃/	in
/œ̃/	un
/ɔ̃/	on
/j/	yy
/ɥ/	uy
/w/	ww

Annexe 10 :

Exemples d'annotations

Exemple 1 : « Listoir du peti cha » (**3129**)

```
<segmentation type="hyposegmentation">
        <segment ecart="orthographique">
                <grapheme ecart="normé" fonction="lexicale" sonorite="consonne" phoneme="ll[10]">L</grapheme>
                <signeGraphique ecart="omis">'</signeGraphique>
        </segment>
        <segment ecart="orthographique">
                <grapheme ecart="omis" fonction="lexicale" sonorite="muet" valeur="h"></grapheme>
                <grapheme ecart="normé" fonction="lexicale" sonorite="voyelle" phoneme="ii">i</grapheme>
                <grapheme ecart="normé" fonction="lexicale" sonorite="consonne" phoneme="ss">s</grapheme>
                <grapheme ecart="normé" fonction="lexicale" sonorite="consonne" phoneme="tt">t</grapheme>
                <grapheme ecart="normé" fonction="lexicale" sonorite="voyelle" phoneme="wwaa">oi</grapheme>
                <grapheme ecart="normé" fonction="lexicale" sonorite="consonne" phoneme="rr">r</grapheme>
                <grapheme ecart="omis" fonction="lexicale" sonorite="muet" valeur="e"></grapheme>
        </segment>
</segmentation>
        <segment ecart="normé">du</segment>
        <segment ecart="orthographique">
                <grapheme ecart="normé" fonction="lexicale" sonorite="consonne" phoneme="pp">p</grapheme>
                <grapheme ecart="normé" fonction="lexicale" sonorite="voyelle" phoneme="ee">e</grapheme>
                <grapheme ecart="normé" fonction="lexicale" sonorite="consonne" phoneme="tt">t</grapheme>
                <grapheme ecart="normé" fonction="lexicale" sonorite="voyelle" phoneme="ii">i</grapheme>
                <grapheme ecart="omis" fonction="lexicale" sonorite="muet" valeur="t"></grapheme>
        </segment>
        <segment ecart="orthographique">
                <grapheme ecart="normé" fonction="lexicale" sonorite="consonne" phoneme="ch">ch</grapheme>
                <grapheme ecart="normé" fonction="lexicale" sonorite="voyelle" phoneme="aa">a</grapheme>
                <grapheme ecart="omis" fonction="lexicale" sonorite="muet" valeur="t"></grapheme>
        </segment>
```

Exemple 2 : « [...] pandans ce sa Maman dor avec c'est chaton [...] » (**1301**)

```
        <segment ecart="orthographique">
                <grapheme ecart="normé" fonction="lexicale" sonorite="consonne" phoneme="pp">p</grapheme>
                <grapheme ecart="orthographique" fonction="lexicale" sonorite="voyelle" phoneme="an" valeur="en">an
                </grapheme>
                <grapheme ecart="normé" fonction="lexicale" sonorite="consonne" phoneme="dd">d</grapheme>
                <grapheme ecart="normé" fonction="lexicale" sonorite="voyelle" phoneme="an">an</grapheme>
                <grapheme ecart="orthographique" fonction="lexicale" sonorite="muet" valeur="t">s</grapheme>
        </segment>
        <segment ecart="phonographique">
                <grapheme ecart="valeur" fonction="lexicale" sonorite="consonne" phoneme="kk" valeur="qu">c
                </grapheme>
                <grapheme ecart="normé" fonction="lexicale" sonorite="voyelle" phoneme="ee">e</grapheme>
```

[10] La forme phonologique est donnée au format LIA développée dans la partie V.

```
</segment>
<segment ecart="normé">sa</segment>
<segment ecart="normé">Maman</segment>
<segment ecart="orthographique">
        <grapheme ecart="normé" fonction="lexicale" sonorite="consonne" phoneme="dd">d</grapheme>
        <grapheme ecart="normé" fonction="lexicale" sonorite="voyelle" phoneme="oo">o</grapheme>
        <grapheme ecart="normé" fonction="lexicale" sonorite="consonne" phoneme="rr">r</grapheme>
        <grapheme ecart="omis" fonction="grammaticale" sonorite="muet" valeur="t"></grapheme>
</segment>
<segment ecart="normé">avec</segment>
<segment ecart="orthographique">
        <segGraph ecart="figure de mot" valeur="ses">c'est</segGraph>
</segment>
<segment ecart="orthographique">
        <grapheme ecart="normé" fonction="lexicale" sonorite="consonne" phoneme="ch">ch</grapheme>
        <grapheme ecart="normé" fonction="lexicale" sonorite="voyelle" phoneme="aa">a</grapheme>
        <grapheme ecart="normé" fonction="lexicale" sonorite="consonne" phoneme="tt">t</grapheme>
        <grapheme ecart="normé" fonction="lexicale" sonorite="voyelle" phoneme="on">on</grapheme>
        <grapheme ecart="omis" fonction="grammaticale" sonorite="muet" valeur="s"></grapheme>
</segment>
```

Annexe 11 :

Module *graph.pm*

```
#!/usr/bin/perl
# Script permettant le découpage en graphème et attribuant une valeur
phonique (ou une absence de valeur phonique) à chaque graphème

package phon;

sub graph {
      my ($mot) = shift(@_);
      my %graph;
      my @graphemes;
      my $c=0;

      if($mot=~/^(\w+)([\r\n])$/){
            $mot=$1;+
            ### Graphèmes digraphes ###
            if($mot=~/(.*)(|^#)ch(.*)/g){ # Graphème ch
                  $mot=$1.$2.$c."#c".$c."#h".$3;
                  my @f=("ch","ch");
                  $graph{$c}=\@f;
                  $c++;
            }if($mot=~/(.*)(|^#)oi(.*)/g){ # Graphème oi
                  $mot=$1.$2.$c."#o".$c."#i".$3;
                  my @f=("oi","wwaa");
                  $graph{$c}=\@f;
                  $c++;
            }if($mot=~/(.*)(|^#)on(.*)/g){ # Graphème on
                  $mot=$1.$2.$c."#o".$c."#n".$3;
                  my @f=("on","on");
                  $graph{$c}=\@f;
                  $c++;
            }if($mot=~/(.*)(|^#)om(.*)/g){ # Graphème om
                  $mot=$1.$2.$c."#o".$c."#m".$3;
                  my @f=("om","on");
                  $graph{$c}=\@f;
                  $c++;
            }

            ### Graphèmes contextuels ###
            ## Initiaux ##
            if($mot=~/^s(.*)$/){ # Graphème s initial
                  $mot=$c."#s".$1;
                  my @f=("s","ss");
                  $graph{$c}=\@f;
                  $c++;
            }
            ## Finaux ##
            if($mot=~/(.*[^#])t(s?)$/){ # Graphème t final
                  $mot=$1.$c."#t".$2;
                  my @f=("t","");
                  $graph{$c}=\@f;
                  $c++;
            }if($mot=~/(.*[^#])s$/){ # Graphème s final
                  $mot=$1.$c."#s";
                  my @f=("s","");
                  $graph{$c}=\@f;
                  $c++;
            }if($mot=~/(.*[^#])e(s?)$/){ # Graphème e final
                  $mot=$1.$c."#e".$2;
```

```
                my @f=("e","");
                $graph{$c}=\@f;
                $c++;
        }

        ### Règles générales ###
        if($mot=~/^(.*[^#]|)a(.*)$/){ # Graphème a
                $mot=$1.$c."#a".$2;
                my @f=("a","aa");
                $graph{$c}=\@f;
                $c++;
        }if($mot=~/^(.*[^#]|)b(.*)$/){ # Graphème b
                $mot=$1.$c."#b".$2;
                my @f=("b","bb");
                $graph{$c}=\@f;
                $c++;
        }if($mot=~/^(.*[^#]|)r(.*)$/){ # Graphème r
                $mot=$1.$c."#r".$2;
                my @f=("r","rr");
                $graph{$c}=\@f;
                $c++;
        }if($mot=~/^(.*[^#]|)s(.*)$/){ # Graphème s
                $mot=$1.$c."#s".$2;
                my @f=("s","zz");
                $graph{$c}=\@f;
                $c++;
        }if($mot=~/^(.*[^#]|)t(.*)$/){ # Graphème t
                $mot=$1.$c."#t".$2;
                my @f=("t","tt");
                $graph{$c}=\@f;
                $c++;
        }

    # Constitution du tableau de résultat contenant les graphèmes de la
forme donnée en entrée
    while ($mot =~/\w+/){
        my @g=@{$graph{substr($mot,0,1)}};
        push(@graphemes,\@g);
        if(substr($mot,0,1) eq substr($mot,3,1)){
            if(substr($mot,0,1) eq substr($mot,6,1)){
                $mot=substr($mot,9,length($mot));
            }else{
                $mot=substr($mot,6,length($mot));
            }
        }else{
            $mot=substr($mot,3,length($mot));
        }

    }
    return @graphemes;
}
1;
```

MOTS-CLÉS : détection d'erreurs, annotation, orthographe, corpus scolaire

RÉSUMÉ

L'intérêt pour l'étude des corpus scolaires, tout en étant grandissant, se heurte à la taille de ces corpus et donc à la difficulté d'une analyse entièrement manuelle. Utiliser des méthodes empruntées au traitement automatique des langues (TAL) pourrait aider à l'exploitation de ces corpus. Cela représente cependant un défi pour le TAL du fait de l'éloignement de ces corpus à la norme. L'objectif de notre travail est d'adapter certaines techniques du TAL, éprouvées par ailleurs, afin de faciliter la constitution et l'exploitation d'un corpus recueilli en classe de CP. L'enjeu est donc double. Il s'agit à la fois de proposer une première définition d'un outil répondant aux besoins de la recherche en linguistique et en didactique. Mais il s'agit également, pour le TAL, de caractériser et de modéliser un type d'écrit distant de la norme. Nous proposerons dans ce mémoire un premier schéma d'annotation d'erreurs et des pistes pour l'analyse automatique de ce type de corpus.

KEYWORDS: spell checking, spelling errors, learner corpora

ABSTRACT

Whereas interest for learner has corpora increased, this research deals with the size of those corpora. Difficulties exist from manual treatments. Therefore we propose to use NLP (Natural Language Processing) methods to help exploit those corpora. This represents a challenge for NLP due to numerous errors from the age level. Our work aims to adapt some verified methods from NLP to build and exploit a first grade elementary school corpus. Our project has two goals in mind. First we hope to construct a framework which can deal with needs in didactic's and linguistic's research. And secondly we aim to model this particular writing type which is far from standard spelling. In this master's thesis we will present a proposition of annotation schema and suggestions for future research.

UNIVERSITE Stendhal GRENOBLE 3

Déclaration anti-plagiat

DECLARATION

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.

Signature :