CORPUS

Corpus

16 | 2017
Spécificités et contraintes des grands corpus de textes scolaires : problèmes de transcription, d'annotation et de traitement

Apports du TAL à la constitution et à l'exploitation d'un corpus scolaire au travers du développement d'un outil d'annotation orthographique

Claire Wolfarth, Claude Ponton et Corinne Totereau

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/2796
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Édition imprimée
Date de publication : 1 janvier 2017
ISBN : 16638-9808
ISSN : 1638-9808

Référence électronique
Claire Wolfarth, Claude Ponton et Corinne Totereau, « Apports du TAL à la constitution et à l'exploitation d'un corpus scolaire au travers du développement d'un outil d'annotation orthographique », *Corpus* [En ligne], 16 | 2017, mis en ligne le 06 janvier 2018, consulté le 08 septembre 2020. URL : http://journals.openedition.org/corpus/2796

Apports du TAL à la constitution et à l'exploitation d'un corpus scolaire au travers du développement d'un outil d'annotation orthographique

Claire Wolfarth, Claude Ponton et Corinne Totereau

1. Introduction

1 Dans un contexte de développement de la linguistique de corpus et des recherches sur l'écriture et les processus d'écriture, alors qu'on n'a jamais autant communiqué par écrit, les écrits des apprenants sont très peu représentés dans les corpus de grande taille constitués ou en cours de constitution. Ceux-ci permettraient pourtant de suivre le processus évolutif à l'œuvre dans l'apprentissage de l'écriture. Écrire un texte est en effet une activité complexe et apprendre à écrire des textes un processus qui s'inscrit dans la durée et qui est loin d'être achevé en fin de scolarité primaire (Fayol, 2013).

2 Quand ils existent, les grands corpus d'apprenants concernent les langues secondes (Granger, 2009 ; Agren, 2004) et sont orientés vers l'analyse des erreurs. Concernant l'anglais langue première, un recueil organisé sur 3 années dans 37 écoles a permis de mettre à disposition le Lancaster Corpus of Children's Project Writing[1]. L'objectif du projet était de faciliter l'exploration de la diversité des écrits des élèves. Le corpus rassemble des textes produits par des enfants de 9 à 11 ans de 1996 à 2000. Le corpus longitudinal complet disponible en ligne concerne 12 élèves seulement (Smith, McEnery, 1998).

3 Dans le domaine du français langue première ou langue de scolarisation, les corpus sont plus limités et difficilement accessibles. Par exemple, le corpus constitué par Christophe Leblay et Emmanuelle Auriac-Slusarcyk (2010) est composé d'écrits scolaires répondant à deux consignes proposées en CE2, CM2, 6e et 4e, ayant pour objectif de faire produire aux élèves un récit et un compte rendu scientifique.

4 Ce sont quelques 500 textes qui ont été rassemblés à l'articulation école-collège sous la direction de Marie-Laure Elalouf en 2005, dans huit classes de CM2 et de 6e : les différents états du texte sont collectés durant une séquence et les liens entre dispositifs didactiques et écrits produits sont analysés dans une perspective génétique (Elalouf, 2005). Les corpus ainsi constitués restent de petite taille (500 textes maximum), relativement hétérogènes et difficilement accessibles.

5 La description des caractéristiques linguistiques des textes scolaires produits en français par des élèves de 6 à 12 ans et de leur évolution reste donc à faire. Elle permettrait d'une part de mettre à la disposition des linguistes des écrits ordinaires d'apprenants, d'autre part de nourrir le travail des didacticiens par la compréhension des dynamiques d'écriture à l'œuvre dans les écrits scolaires. Ce travail permettrait par ailleurs de soutenir, par la constitution d'une banque de textes accessible à tous les professeurs, l'enseignement de l'écriture à l'école, dont la plupart des observateurs s'accordent à dire qu'il est insuffisant, et ce dès le cours préparatoire (Bouysse, 2006, p. 10).

6 Dans ce contexte, nous travaillons à la collecte et à l'édition d'un grand corpus numérique longitudinal de textes narratifs scolaires. L'objectif de ce projet est de réaliser une description linguistique des structures utilisées par les élèves au cours de la construction de leurs apprentissages de l'écrit (morphographie, syntaxe, lexique, structuration du discours), ainsi que de l'évolution des procédés d'écriture à différents moments de la scolarisation à l'école primaire.

7 À terme, le corpus devrait contenir plusieurs milliers de productions. Un tel corpus ne pouvant être finement analysé manuellement, son exploitation sera facilitée par un module d'annotation empruntant des méthodes au traitement automatique des langues (TAL). Ce module devrait permettre une aide automatique à l'annotation de nombreux phénomènes linguistiques (orthographiques, syntaxiques, lexicaux, etc.), permettant une grande variété d'utilisation du corpus. Par ailleurs, le recours au TAL devrait permettre à terme une interrogation fine du corpus par les chercheurs et les enseignants.

8 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 adoptons 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 (production d'élèves de CP selon une consigne connue) pour une aide à l'analyse. Nous nous plaçons donc résolument dans une perspective d'aide à l'exploitation du corpus et non pas dans une perspective de détection et de diagnostic entièrement automatique. Dans une approche empirique de cette problématique, nous nous focalisons ici sur la détection et l'annotation des erreurs d'orthographe (section 3). Avant cela, nous présentons les spécificités de notre corpus, en termes de recueil et de numérisation (section 2).

2. Le projet Scoledit

9 Le projet Scoledit a pour objectif de procéder à la collecte, à l'annotation et à l'édition d'un grand corpus numérique longitudinal de textes narratifs et descriptifs. L'enjeu scientifique du projet est de permettre de rendre compte des évolutions des procédés

d'écriture à différents moments de la scolarisation de l'école primaire en rassemblant plus de trois mille textes rédigés du CP au CM2 à partir d'un protocole commun. Ce projet prend appui sur le recueil de données effectué dans le cadre du projet national « Lire - Écrire au CP », 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). Dans le cadre de ce projet, un premier recueil, en juin 2014, a permis de rassembler 2507 productions provenant de 131 classes de CP. En juin 2015, un second recueil a eu lieu auprès des mêmes élèves alors en classe de CE1, qui a permis de rassembler 2049 textes. Par la suite, le travail portera sur les productions de 57 classes, réparties dans cinq académies (Bordeaux, Grenoble, Lyon, Toulouse, Clermont-Ferrand), parmi les 131 précédemment citées. Le recueil se prolongera chaque fin d'année scolaire jusqu'en 2018, année où les élèves seront alors en CM2.

10 Dans la suite de cet article, nous nous concentrerons exclusivement sur les 1 169 productions recueillies dans les classes de CP des 5 académies retenues, seules productions disponibles au moment de l'élaboration de l'outil présenté.

2.1 Le recueil du corpus

11 Cette section présente la méthode utilisée pour recueillir les productions de CP de notre corpus[2]. Lors de la phase de collecte, quatre images (figure 1) étaient présentées 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 ». Lors de la rédaction, les élèves pouvaient, au besoin, consulter les images au tableau ou au dos de leur feuille.

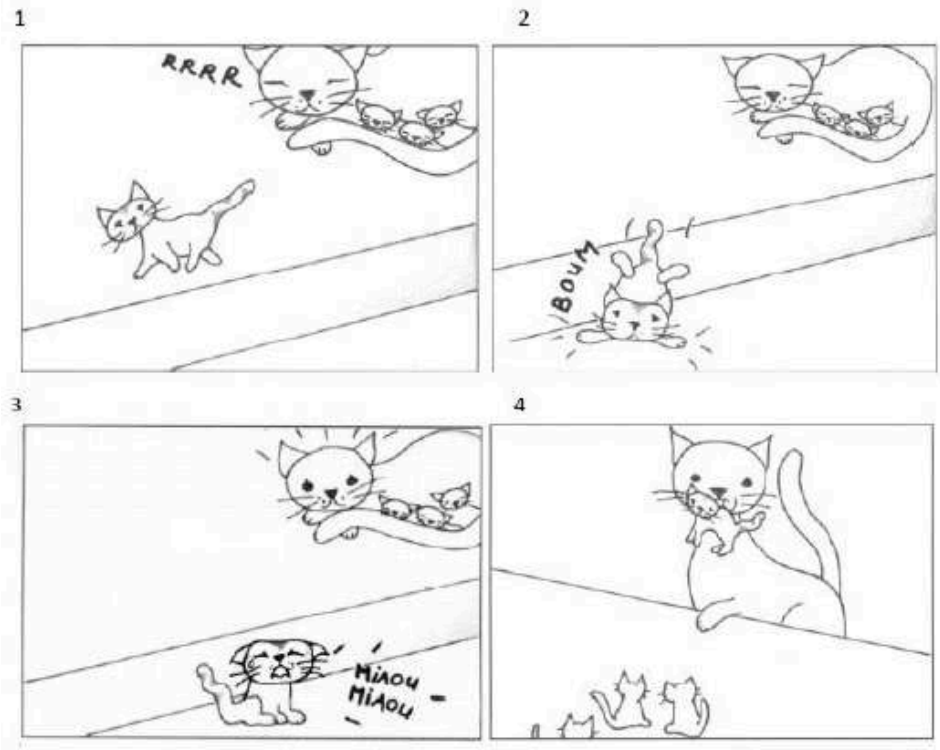


Figure 1. Images présentées aux élèves
lors de la production écrite en CP

12 Au niveau national, ce procédé a donné lieu à 2 507 productions relativement longues au regard de ce qui est attendu d'un élève de fin de CP (Goigoux, 2016). En effet, plus de 85 % des élèves écrivent plus de 60 caractères et 30 % d'entre eux en produisent plus de 100. Précisons également que ce n'est qu'une très petite minorité d'élèves qui ne produit rien puisqu'ils sont moins de 1 %.

13 Dans notre corpus de 907 productions[3], on retrouve les mêmes tendances. En effet, 90 % des élèves ont produit plus de 60 lettres et 31 % plus de 100. De même, les élèves n'ayant rien produit représentent moins de 1 % et la moyenne se situe aux alentours de 83 lettres produites ; la production la plus longue étant de 248 lettres.

14 En termes de mots, la moyenne de notre corpus se situe aux alentours de 23 mots par productions. La longueur des productions s'étend de 0 à 62 mots, avec une exception à 92 mots.

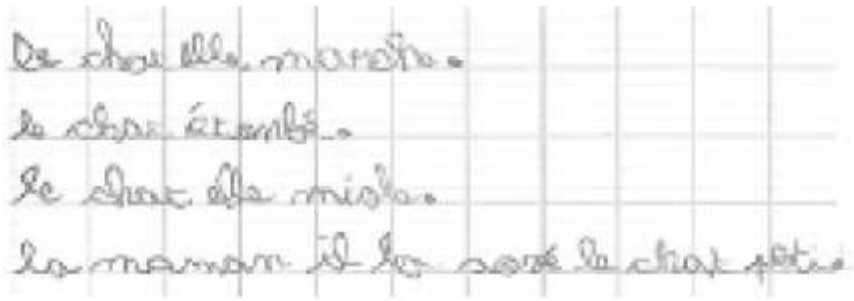
2.2 La numérisation du corpus

15 Afin de le rendre accessible et d'en permettre un traitement informatique, le corpus a été numérisé ; chaque production a été scannée, puis transcrite. Des conventions de transcription ont alors été élaborées, permettant de rendre cette transcription homogène sur l'ensemble du corpus, prérequis à toute analyse automatique.

16 La transcription du corpus a été pensée pour une utilisation par les linguistes, didacticiens et enseignants et non par des généticiens du texte par exemple. Cette perspective a influencé le choix des phénomènes à annoter. Il a été décidé de privilégier le texte final plutôt que sa genèse ou ses caractéristiques visuelles.

17 Ainsi, certains éléments comme les dessins, les traces de gomme, les ratures ou les autres traces de révision ont été simplement marqués de manière très générique avec des balises du type <revision/> ou <dessin/> sans détail supplémentaire[4]. Il sera donc possible ultérieurement de retrouver ces productions pour affiner si besoin la description. À ce sujet, un groupe national, auquel nous participons, travaille à la mise au point d'un système de conventions communes.

18 Nos conventions de descriptions conservent toutefois certaines marques visuelles porteuses de sens. C'est le cas de certains retours à la ligne. Les signes // marquent les retours à la ligne perçus comme volontaires, pouvant référer à une réalité sémantique ou syntaxique. Le signe / marque les retours à la ligne contraints spatialement par l'extrémité de la feuille.



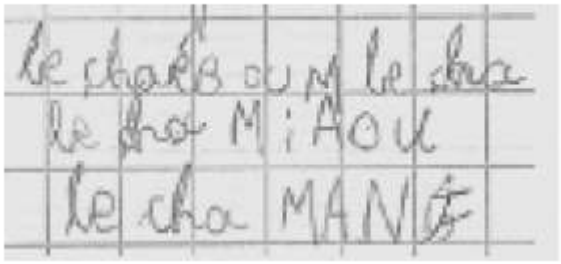
19 « le chat elle marche. // le chat étonbé. // le chat elle miole. // la maman il la sové le chat peti. » (1361, Le chat, il marche. Le chat est tombé. Le chat, il miaule. La maman, elle a sauvé le petit chat.)

20 Le reste de nos conventions porte sur les éléments d'incertitude lors de la transcription. En effet, certains caractères peuvent être ambigus ou difficilement déchiffrables[5]. Les choix qui sont faits lors de la transcription de ces passages auront une influence directe sur la suite du traitement du corpus, c'est pourquoi il nous semble important de les relever.

21 La balise **<letMF>x</letMF>** permet d'annoter des signes graphiques ne correspondant à aucune lettre du système graphique du français mais qui peuvent être interprétés grâce au contexte. Dans l'exemple suivant, le signe peut être interprété comme la lettre *i* ou la lettre *t*. Toutefois, dans le contexte ci-dessous, il est identifiable de manière formelle comme une lettre *t* mal formée :

22 « 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 [...])

23 Lorsque le contexte ne permet pas de désambigüiser un signe de manière aussi évidente, la balise **<x|y>** permet d'inclure les différentes possibilités dans la transcription. L'interprétation la plus probable est mise en avant.



24 « [...] le chaéB OUM le cha // le cha MIAOU // le cha MAN<J|G> » (2930, Le chat et BOUM le chat. Le chat MIAOU. Le chat mange.)

25 Quelles que soit les consignes de transcription, ce travail manuel d'interprétation est sujet à erreurs : approches différentes des transcripteurs (*i. e.* choix de *J* ou de *G* comme lettre la plus plausible dans l'exemple précédent), erreurs humaines de saisie, etc. Toutefois, le schéma de transcription tel qu'il est défini actuellement répond à nos objectifs sans pour autant alourdir le temps de transcription (de l'ordre d'une minute trente environ par production de CP pour 65 caractères en moyenne) et a permis une annotation relativement consensuelle entre les transcripteurs.

2.3 La mise à disposition du corpus

26 L'ensemble du corpus sera mis en ligne afin d'en permettre une large diffusion auprès des enseignants et des chercheurs qui désirent le consulter. Pour chaque production, le scan et la transcription seront disponibles. Il sera alors possible de parcourir les productions par niveau scolaire ou de manière longitudinale, par élèves. Dans une perspective d'amélioration de la qualité de nos données de transcription, les utilisateurs de ce site pourront laisser un commentaire sur chaque production.

27 Au vu de la taille du corpus final, ces productions ne peuvent être que difficilement exploitées manuellement par les didacticiens et les enseignants qui en auront l'usage. C'est pourquoi, nous proposons d'accompagner ce corpus d'un outil facilitant son exploration. Cet outil consiste en un module d'interrogation du corpus. Nous prévoyons ainsi de permettre des réponses à des questions du type : à quel moment de l'apprentissage apparaissent les marques du pluriel non audibles ? Comment évoluent les marques de conjugaison à l'imparfait ? Quelles sont les erreurs orthographiques les plus fréquentes au CM1 ? Pour répondre à ces questions, une annotation des phénomènes linguistiques concernés est nécessaire.

3. Le module d'annotation des erreurs

28 Du choix des phénomènes à annoter et de la façon dont ils le seront dépendent directement les possibilités de requêtes et d'utilisation de ce corpus. Nous présenterons ici une étude centrée sur l'annotation des erreurs orthographiques, l'orthographe étant un des thèmes majeurs de recherche des porteurs du projet.

3.1 Revue des méthodes de détection et de correction d'erreurs

29 Il existe en TAL de nombreux travaux s'intéressant à la détection et la correction d'erreurs, que ce soit dans le domaine des vérificateurs d'orthographe (Antidote[6], Cordial[7]...) ou dans le domaine de la normalisation. Dans ce dernier cas, il s'agit généralement d'un prétraitement avant analyse d'un contenu (particulièrement développé pour les données provenant du web, des réseaux sociaux ou encore des SMS).

30 Les méthodes utilisées sont très diverses selon la nature des corpus en présence. Dans le cadre des correcteurs orthographiques « classiques », destinés aux scripteurs experts, les erreurs prises en compte sont généralement des erreurs de performance (erreurs de saisie au clavier). Dans ce cas, il est souvent possible de procéder à des comparaisons graphiques avec un lexique de formes normées (Damerau, 1964 ; Kernighan, Church et Gale, 1990). Il est également possible de procéder à des corrections basées sur des approches contextuelles (Brill et Moore, 2000 ; Carlson et Fette, 2007 ; Park et Levy, 2011). Cependant, ces approches sont peu efficaces pour les textes très fautifs.

31 Des systèmes se sont également confrontés aux textes peu normés, parmi lesquels les corpus d'apprenants en FLE et les corpus récoltés via les réseaux sociaux ou les SMS. Toutefois, dans le cadre du traitement automatique des corpus de FLE, beaucoup de travaux se sont essentiellement centrés sur des systèmes de corrections en contexte restreint, où les réponses sont contraintes par les questions, pour lesquelles il est possible de se baser sur une comparaison avec la réponse attendue (Chanier, 1996), sur des listes d'erreurs non natives (Mitton 1996) et sur des systèmes de règles applicables élaborées à la main (Chanier, 1992).

32 Dans le cadre de la normalisation de messages provenant de tweets ou de SMS, différentes approches ont été proposées. Certains systèmes s'appuient notamment sur des corrections systématiques (Baranes, 2012), ce qui permet de gérer les phénomènes d'abréviations, etc. D'autres se basent sur la réalité phonémique du texte pour retrouver la forme normée (Beaufort *et al.*, 2010). Cette dernière approche va particulièrement nous intéresser pour la suite de notre travail.

3.2 Précisions méthodologiques

33 Le module d'annotation des erreurs d'orthographe s'appuie sur des méthodes et des outils issus du traitement automatique des langues. La plupart de ces méthodes prennent comme unité de traitement le mot, souvent défini comme une suite de caractères séparées par des espaces ou de la ponctuation (à l'exception de certains signes comme les tirets). Dans les écrits d'apprenants où la segmentation en mots n'est pas encore maitrisée par tous, cette définition ne correspond pas toujours à des mots. On constate en effet des cas d'hypersegmentation, où l'enfant divise une unité lexicale en plusieurs formes graphiques, exemple *après*, orthographié *à prè* (1156). À l'inverse, on rencontre également des cas d'hyposegmentation, où l'enfant agglomère plusieurs unités entre elles, comme *l'attrape* écrit *latrape* (1116)[8]. Dans la suite, nous parlerons de « formes » au sens large pour désigner ces unités.

34 Pour annoter les erreurs d'orthographe, nous nous référons à la fois à la théorie de Catach (1979, 1980, 1995) et à celle de Blanche-Benveniste et Chervel (1969, 1978), reprise par Cappeau et Roubaud (2005). Nous reprenons à Nina Catach sa définition du

graphème comme « plus petite unité distinctive et/ou significative de la chaine écrite » (Catach, 1980, p. 16) et son organisation en plurisystème. Nous utiliserons ainsi les termes de phonogrammes (« graphème susceptible d'avoir un correspondant phonique », Catach, 1979, p. 27) et de logogrammes (signes graphiques permettant de distinguer les homophones les plus courants). À l'exemple de Jaffré (Fayol et Jaffré, 2008), nous distinguons la phonographie (qui concerne la transcription des phonèmes des unités lexicales) de la sémiographie (qui concerne le sens des mots)[9].

35 En revanche, en ce qui concerne la typologie des erreurs, nous nous baserons sur la classification de Cappeau et Roubaud qui distingue les erreurs de code phonographique des erreurs de sélection de la norme orthographique. Les premières correspondent aux erreurs effectuées lors de la transcription phonographique et implique une modification de la représentation phonologique du terme encodé (par exemple *tondé*, forme normée : *tombé*). Les secondes désignent les erreurs commises lors de la sélection des différentes graphies possibles pour une même représentation phonologique (par exemple *toudincou*, forme normée : *tout d'un coup*).

36 Pour repérer et annoter une erreur d'orthographe, différentes étapes sont nécessaires : (1) la détection des formes (cf. § 3.2) comportant une ou plusieurs erreurs ; (2) l'identification de la forme attendue ; (3) l'annotation de l'erreur ou des erreurs. Dans cet article, nous présenterons chacune de ces étapes tout en nous concentrant particulièrement sur les formes présentant des erreurs de sélection de la norme orthographique. La graphie de ces formes transcrit de manière exacte la forme phonologique mais ne respecte pas la norme orthographique.

37 La suite de ce travail nécessitant une première analyse manuelle, un échantillon d'étude de 20 productions a été sélectionné, soit 471 formes. Nous pensons que cet échantillon est suffisant pour une première approche dont le but est uniquement de dégager de grandes tendances.

3.3 Reconnaissance des formes erronées

38 Le travail préliminaire à notre analyse consiste à repérer les erreurs orthographiques présentes dans notre corpus. Il doit permettre d'identifier les formes concernées par le travail d'annotation.

39 Pour ce travail, nous considérons comme erreurs orthographiques toute forme non contenue dans un lexique de formes fléchies du français donné (celui de TreeTagger dans notre cas). Cette hypothèse se justifie par une très faible présence de noms propres et l'inexistence de termes spécialisés dans nos productions de CP, peu de formes devraient donc être considérées comme erronées à tort. En revanche, elle laisse de côté de nombreuses erreurs, notamment les erreurs non lexicales comme *chat* (52, forme attendue *chats* dans « [...] 5 petit chat. »).

40 L'emploi de TreeTagger (Schmid, 1994) dans notre module permet d'appliquer ce principe à notre corpus. TreeTagger est un étiqueteur morphosyntaxique basé sur un lexique propre de formes fléchies du français. Il permet de segmenter un texte en formes et d'attribuer à chaque forme une catégorie grammaticale et un lemme. Lorsque la forme n'est pas une forme connue dans son lexique, elle est étiquetée <unknown>. Ainsi, la forme *maman* est étiquetée : [maman, NOM, maman], tandis que la forme *fraire* est étiquetée : [fraire, NOM, <unknown>].

41 Comme nous l'avons évoqué précédemment, l'emploi de cet outil pour identifier les erreurs pose certains problèmes. Dans les paragraphes suivants, nous tenterons une analyse des résultats donnés par TreeTagger, non dans un but de typologie d'erreurs, mais uniquement pour identifier les problèmes sous-jacents à notre hypothèse de travail et sur lesquels il nous faudra revenir dans un travail futur.

42 Sur un échantillon de 471 formes, TreeTagger a attribué une catégorie grammaticale et un lemme à 353 d'entre elles. Ces formes sont donc, selon notre hypothèse, considérées comme non erronées. Or, parmi ces 353 formes reconnues, 51 comportaient des erreurs. Ces erreurs non détectées sont principalement :

> Des erreurs logographiques (remplacement d'une forme par une forme homophone), comme *et* (1556, forme attendue *est*) dans « [...] La maman et venus [...] » ;
> Des erreurs morphographiques, comme *chaton* (1336, forme attendue *chatons*) dans « les chaton [...] » ;
> Des erreurs de segmentation, comme *des sendu* (1336, forme attendue *descendu*).

43 En revanche, 118 formes n'ont pas pu être lemmatisées. Elles sont donc, toujours selon notre hypothèse, considérées comme erronées. Parmi ces 118, 114 d'entre elles comportent effectivement une ou plusieurs erreurs, mais 4 n'en comportent aucune. Il s'agit de l'onomatopée *r r r r* recopiée par certains élèves et identifiée comme 4 formes distinctes inconnues pour TreeTagger. Ce cas étant relativement marginal dans l'ensemble du corpus, nous pouvons considérer les formes étiquetées <unknown> par TreeTagger comme porteuses d'erreurs.

44 Pour résumer, notre hypothèse de départ ne permet pas d'identifier toutes les erreurs de notre corpus, cependant les formes retenues contiennent des erreurs dans une forte probabilité. Cette hypothèse nous permet donc d'obtenir une base pour un premier travail exploratoire non exhaustif. Nous appellerons donc désormais forme erronée, toute forme inconnue de TreeTagger.

3.4 Identification de la forme attendue

45 Une fois les formes comportant des erreurs détectées, il va nous falloir identifier les erreurs afin de les annoter. L'identification et l'annotation des erreurs se font par comparaison de la forme erronée trouvée en corpus et de la forme attendue, il est donc nécessaire de déterminer la forme attendue au préalable.

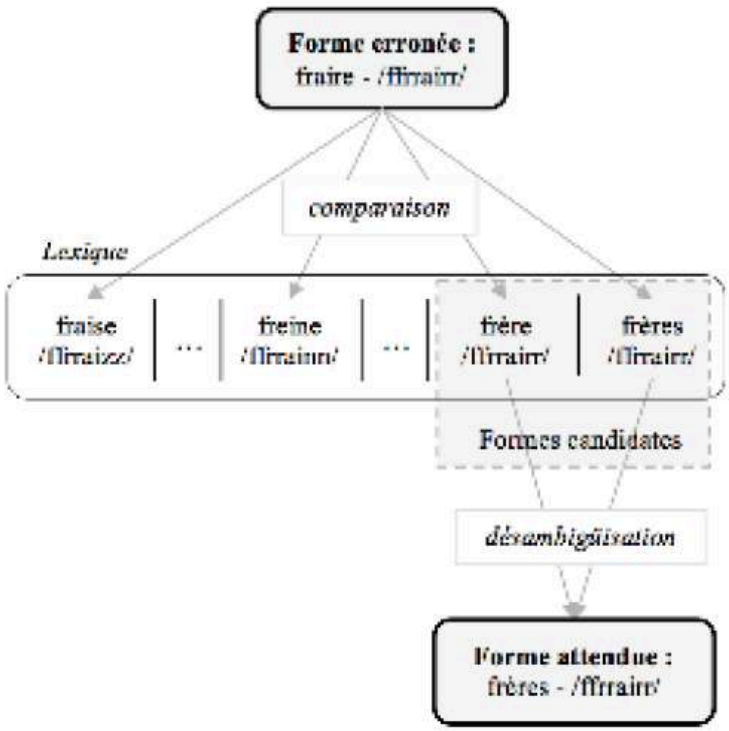
46 Dans cette partie, nous ne décrirons que le processus envisagé pour déterminer la forme attendue des formes présentant des erreurs de *sélection de la norme orthographique* telles que nous les avons présentées précédemment (cf. 3.2). Pour rappel, il s'agit de formes pour lesquelles la forme phonologique est conservée mais ne respectant pas la norme orthographique du français. L'échantillon utilisé ne comportant pas suffisamment d'exemples de ce type d'erreur, il a été doublé. Les mesures suivantes se basent donc sur un nouvel échantillon de 40 productions, soit 936 formes.

47 À l'instar des méthodes éprouvées dans certains travaux de correction automatique ou de normalisation de SMS (Laporte et Silberztein, 1989 ; Williams et Maier, 1991 ; Beaufort *et al.*, 2010), nous proposons ici un passage par la représentation phonologique pour identifier la forme normée attendue. En effet, pour ce type d'erreurs, la

représentation phonologique de la forme en contexte est la même que la représentation phonologique de la forme attendue, ce qui constitue un indice fort pour retrouver cette forme.

48 Comme le montre le schéma suivant (figure 2), le résultat de cette comparaison est l'extraction d'une ou de plusieurs formes dont la représentation phonologique est identique à celle de la forme erronée, ce sont les formes candidates. L'hypothèse est que la forme normée se trouve parmi ces formes candidates. L'étape qui permet de choisir cette forme est appelée étape de désambigüisation.

Figure 2. Procédure d'identification de la forme attendue



49 Cependant, pour que l'étape de désambigüisation soit optimale, il est nécessaire de limiter le nombre de formes candidates en limitant la taille du lexique utilisé pour la comparaison.

50 Par ailleurs, si la forme attendue est présente dans la liste des formes candidates, on considèrera que la forme attendue est identifiée (FAI) et il sera possible de procéder à l'étape de dés-ambiguïsation. Dans le cas où elle ne l'est pas, on considèrera qu'elle est non identifiée (FANI) et la désambiguïsation ne peut qu'échouer.

51 L'enjeu est donc de déterminer le lexique permettant d'identifier la forme attendue tout en limitant le nombre de formes candidates. Nous faisons alors l'hypothèse que l'utilisation de la connaissance du contexte de production dans le choix de notre lexique devrait permettre de répondre au mieux à cette double contrainte. Pour ce faire, différentes ressources ont été testées :

> (1) Une liste de formes phonologiques élaborée directement à partir du corpus en cours d'élaboration, il s'agit donc d'une ressource au plus près du contexte de production

(2) Les listes Manulex (Lété, Sprenger-Charolles et Colé, 2004), produites en milieu scolaire, il s'agit donc d'un contexte plus élargi
(3) Lexique 3 (New et Pallier, 2005), lexique des formes fléchies du français, qui vise une certaine exhaustivité et qui se détache donc du contexte de production

52 Dans le travail présenté ici, seule l'étape de comparaison a été étudiée.

3.4.1 Identification de la forme attendue à partir du corpus

53 Nous partons de l'hypothèse que la méthode de recueil du corpus de CP étant fortement contrainte par les images présentées aux scripteurs débutants, ces contraintes ont influencé les écrits des apprenants de telle sorte que le lexique utilisé semble, à première vue assez restreint et redondant. Il apparait donc pertinent d'utiliser cette redondance pour identifier les formes erronées. C'est pourquoi, une première liste (*liste A*) de formes phonologiques a été construite à partir des formes du corpus jugées normées (c'est-à-dire les formes connues de TreeTagger, voir 3.3). Les formes phonétiques correspondantes ont été obtenues à l'aide de LIA_PHON (Bechet, 2001).

54 À l'issue de cette première expérimentation, il est rapidement apparu que certaines formes n'apparaissaient jamais correctement orthographiées en corpus (*i. e.* la forme *dort* orthographiée tantôt *dor* : 95, 130, 145... tantôt *dore* : 652, 660, 807... mais jamais *dort*[10]). Une deuxième liste a donc été élaborée (*liste B*) contenant l'ensemble des formes du paradigme flexionnel des lemmes des formes de la *liste A*.

55 Après phonétisation, cette ressource lexicale a été testée sur notre échantillon de 40 productions présentant 77 formes répondant à nos critères, c'est-à-dire ne contenant que des erreurs ne modifiant pas la représentation phonologique. Une comparaison a donc été faite entre les représentations phonologiques des formes visées et les représentations de la ressource lexicale. Chaque comparaison a donné lieu à une liste de formes candidates. L'enjeu est de savoir si la forme normée attendue, ou tout du moins une des formes du lemme attendu, est incluse dans cette liste (voir 3.4).

56 Les résultats sont résumés dans le tableau ci-après :

Tableau 1. Comparaison du nombre de formes identifiées des lexiques tirés du corpus

	Liste A	**Liste B**
Formes attendues identifiées (FAI)	56	61
Formes attendues non identifiées (FANI)	3	1
Aucune forme normée trouvée (AFN)	18	15
Taux de réussite (FAI / (FAI+FANI+AFN))	72.7 %	79.2 %
Taux de confiance (FAI / (FAI+FANI))	94.9 %	98.4 %
Degré d'ambigüité	1.6	2.5

57 Il nous faut donc envisager 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 (réduire l'ambigüité). Pour nous aider à comparer les différentes méthodes, nous employons trois mesures :

> le calcul du **taux de réussite** : le nombre de fois que la forme attendue est trouvée par rapport au total des formes à identifier ;
> le calcul du **taux de confiance** : permet de déterminer la confiance que nous pouvons avoir dans les listes de formes candidates. Il s'agit du nombre de formes pour lesquelles la forme attendue a été trouvée parmi les formes pour lesquelles au moins une forme normée a été trouvée ;
> **le degré d'ambigüité** : nombre moyen de formes trouvées à l'exclusion des cas où aucune forme n'a été trouvée.

58 Un taux de confiance différent de 100 % signifie que la recherche des formes normées ne permettra pas de trouver toutes les formes attendues alors même que des formes normées sont trouvées. Cela revient donc à inclure un degré d'erreur dès cette étape. Ces listes ne sont donc pas satifaisantes et il nous faut nous tourner vers des listes plus exhaustives.

3.4.2 Identification de la forme attendue à partir du contexte scolaire

59 Malgré une grande redondance du lexique utilisé, certains lemmes ne sont utilisés que dans un très petit nombre de productions et le corpus n'en contient donc pas toujours une forme bien orthographiée. Afin de traiter ces cas précis, un recours à une liste extérieure au corpus est nécessaire. S'agissant d'un corpus de textes scolaires, nous pouvions nous attendre à retrouver un lexique que côtoient couramment les enfants. Nous proposons alors d'utiliser la ressource Manulex. Cette ressource fait état de toutes les formes rencontrées dans une liste donnée de 54 manuels scolaires, selon que l'on s'intéresse aux manuels de CP, de CE1 ou de cycle 3. En outre, elle présente l'avantage de contenir des formes fléchies et non des lemmes, contrairement aux nombreuses autres ressources lexicales scolaires. Pour notre étude, nous nous sommes intéressés aux formes issues du niveau CP (liste *Manulex CP*, 11 332 formes) et de tous les niveaux (liste *Manulex CP-CM2*, 48 887 formes), phonétisées à l'aide de LIA_PHON.

Tableau 2. Comparaison du nombre de formes identifiées par les listes Manulex

	Manulex CP	Manulex CP-CM2
Formes attendues identifiées (FAI)	74	77
Formes attendues non identifiées (FANI)	0	0
Aucune forme normée trouvée (AFN)	3	0
Taux de réussite (FAI / (FAI+FANI+AFN))	96.1 %	100 %
Taux de confiance (FAI / (FAI+FANI))	100 %	100 %
Degré d'ambigüité	2.4	3.2

3.4.3 Identification de la forme attendue à partir d'un lexique général

60 Afin de vérifier la pertinence de l'utilisation d'une ressource spécialisée, correspondant au contexte de recueil du corpus, nous proposons de comparer nos résultats à ceux obtenus avec une ressource lexicale plus exhaustive constituée à partir de ressources orales de locuteurs adultes natifs. Pour ce faire, nous utilisons *lexique 3*, ressource de près de 130.000 formes fléchies[11]. Il est donc attendu que le degré d'ambiguïté soit plus élevé en utilisant cette ressource lexicale. Les résultats sont reportés dans le tableau suivant :

Tableau 3. Comparaison du nombre de formes identifiées sur l'ensemble des listes

	Liste A	Liste B	Manulex CP	Manulex CP-CM2	Lexique 3
Formes attendues identifiées (FAI)	56	61	74	77	77
Formes attendues non identifiées (FANI)	3	1	0	0	0
Aucune forme normée trouvée (AFN)	18	15	3	0	0
Taux de réussite (FAI / (FAI+FANI+AFN))	72.7 %	79.2 %	96.1 %	100 %	100 %
Taux de confiance (FAI / (FAI+FANI))	94.9 %	98.4 %	100 %	100 %	100 %
Degré d'ambigüité	1.6	2.5	2.4	3.2	3.7
Total	**77**				

61 À travers ces tableaux, il apparait qu'utiliser un lexique non spécialisé augmente le degré d'ambigüité sans modifier le taux de réussite et le taux de confiance. 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.

62 Tant la ressource Manulex CP que la ressource Manulex CP-CM2 présentent un taux de confiance de 100 %, ce qui signifie que lorsqu'une ou plusieurs formes normées sont trouvées, la forme attendue est incluse dans les résultats, ces deux ressources sont donc des ressources que nous pourrons utiliser dans la suite de nos travaux. Toutefois, certaines formes utilisées par les scripteurs de notre corpus ne sont pas contenues dans la liste Manulex CP, le taux de réussite n'est en effet que de 96,1 %. Manulex CP-CM2 permet de pallier ce problème mais augmente par là même le degré d'ambigüité. Afin de tirer avantage de ces deux listes, il est intéressant d'utiliser dans un premier temps la liste Manulex CP et, lorsqu'aucune forme normée n'a été trouvée, d'utiliser Manulex CP-CM2 dans un second temps. On obtient alors les résultats suivants :

Tableau 4. Résultats des méthodes par combinaison de listes

Mesure	**Manulex CP + Manulex CP-CM2**
Taux de réussite (FAI / (FAI+FANI+AFN))	100 %
Taux de confiance (FAI / (FAI+FANI))	100 %
Degré d'ambigüité	2.4

63 Le résultat de la comparaison réalisée à l'étape précédente est une liste de formes normées, parmi lesquelles se trouve la forme attendue. Une étape de désambiguïsation est alors nécessaire pour identifier cette forme. Si un travail plus soutenu reste à faire, l'on peut d'ores et déjà avancer que la désambiguïsation basée sur les fréquences semble être efficace pour la plupart des formes erronées, à l'exception des logogrammes fréquents, comme *a/à* et *et/est*.

3.5 Annotation de l'erreur

64 La forme attendue d'une forme erronée étant identifiée, il est alors possible de diagnostiquer et d'annoter l'erreur ou les erreurs contenues par cette forme. L'élément central de cette annotation est la mise en place d'un modèle permettant 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. Ce modèle doit remplir deux conditions :

> être **descriptif** : il doit permettre une description linguistique des productions d'élèves répondant aux besoins des chercheurs ;
> être **opératoire** : il doit permettre une aide automatique efficiente à l'annotation des erreurs en fonction des possibilités du TAL.

65 Mais réaliser un tel schéma soulève de nombreuses questions, à commencer par : quelles erreurs annotons-nous et surtout jusqu'à quel niveau de description ? Quelle méthodologie employons-nous ?

66 L'outil d'aide à l'annotation que nous développons contiendra différents modules selon les niveaux de traitement envisagés. Nous présenterons ici le travail effectué sur l'annotation des erreurs d'orthographe avant de présenter les méthodes automatiques envisagées pour l'annotation des erreurs de type orthographique.

3.5.1 Élaboration d'un modèle d'annotation

67 Ce modèle d'annotation a été élaboré de manière empirique, à partir de l'observation d'un petit nombre de productions. Après relevé des différentes erreurs, celles-ci ont été classées. Ce premier travail de classification s'est inspiré des travaux de Damerau et Levenshtein (Damerau, 1964), qui, dans un travail connu sous le nom de distance de Levenshtein, postulait quatre types d'erreurs : l'omission, l'addition, la substitution et le déplacement d'un ou plusieurs caractères.

68 Puis, il est apparu que, s'agissant d'une écriture liée à une réalité phonique, certaines erreurs pouvaient avoir plus d'impact que d'autres sur cette réalité sonore à l'intérieur

d'une même catégorie. En effet, l'omission du *t* de $tombe_{\text{VERBE}}$ aura un impact plus conséquent sur la réalisation sonore de cette forme que l'omission du *t* dans $chat_{\text{NOM}}$. Or l'importance de cet impact influe sur la fréquence de l'erreur. Ainsi, alors que pour 75 occurrences de la forme *chat*, le graphème *t* est omis 10 fois, il n'est omis dans aucune des 29 formes du verbe TOMBER. Cette différence a donc été incluse à notre classification, rejoignant en cela celle proposée par Nina Catach (1980).

69 En revanche, suite à ce travail d'observation, il ne nous a pas paru pertinent de mener un travail morphologique sur les erreurs. Il semble, en effet, que les élèves de CP se basent plus sur des indices phonologiques que morphologiques.

70 Hormis le déplacement, toutes ces erreurs peuvent être annotées au niveau du graphème. Il s'agit donc des lettres codant les mêmes phonèmes ou ayant la même fonction sémiographique. Pour chaque graphème seront donc données les informations suivantes :

> La réalité phonique du graphème : s'il s'agit d'une voyelle, d'une consonne, d'une semi-voyelle, d'un graphème muet, d'une voyelle auxiliaire[12] (*e* de *rév<u>e</u>ille*) ou d'un graphème muet auxiliaire (*e* de *nage<u>a</u>*) ;
> La fonction : s'il s'agit d'un graphème lexical (i. e. *t* dans *chats*) ou grammatical (i. e. *s* dans *chats*), à l'exception des lettres non attendues encodant la liaison (*des <u>s</u>otres*, forme normée : *des autres*) qui sont considérées comme un graphème pour lequel la fonction donnée sera de transcrire la liaison.
> À ces informations s'ajoute l'objet de cette annotation, c'est-à-dire la nature de l'erreur. Un graphème peut être :
> Normé : lorsqu'il est identique au graphème attendu (exemple : *ch* dans *chat*) ;
> Omis : lorsqu'un graphème est attendu (exemple : *t* dans *cha*, forme normée : *chat*) ;
> Inséré : lorsqu'aucun graphème n'est attendu (exemple : *e* dans *peleure*, forme normée : *pleure*) ;
> Substitué : lorsqu'un graphème est remplacé par un autre. Quatre cas de substitution seront distingués :

71 - L'erreur de sélection de la norme orthographique : lorsque le graphème attendu est remplacé par un graphème transcrivant le même phonème, mais ne correspondant pas à l'orthographe attendue (exemple : *on* dans *t<u>on</u>be*, forme normée : *tombe*) ;

72 - L'erreur de phonologie proche : lorsque le graphème est remplacé par un graphème transcrivant un phonème partageant de nombreux traits phoniques communs au phonème attendu. Ainsi, dans l'exemple *in* [œ̃] dans *d<u>in</u> cou* [ɛ̃], forme normée : *d'un coup* seul le trait arrondi / non-arrondi distingue ces deux phonèmes dont l'opposition est souvent neutralisée (Walter, 1976), de même dans l'exemple *e* [œ] dans *pl<u>e</u>re* [ø], forme normée : *pleure*, distingué par le degré d'aperture, cette opposition tend également à se neutraliser ;

73 - L'erreur de valeur en contexte : lorsque le graphème est remplacé par un graphème pouvant transcrire le même phonème, dans un contexte différent (exemple : *c* dans *<u>c</u>e*, forme normée : *que*) ;

74 - L'erreur de type phonographique : lorsque le graphème est remplacé par un graphème ne transcrivant pas le même phonème, il y a méconnaissance du code phonographique (exemple : *m* dans *maim*, forme normée : *mais*).

75 Toutefois, certaines erreurs ne peuvent s'appréhender au niveau du graphème. En effet, elles ne peuvent être décrites qu'en prenant en compte plusieurs graphèmes à la fois. C'est souvent le cas du déplacement correspondant plutôt à des erreurs d'inversion dans notre corpus, mais également des erreurs logogrammiques (Catach, 1995) (remplacement d'un logogramme par un autre) ou encore des erreurs que nous avons appelées substitution flexionnelle, comme *fit* écrit *fesa* (seules erreurs de type morphologique relevées dans notre annotation). Pour ces erreurs, l'annotation portera sur l'ensemble des graphèmes concernés et seule la nature de l'erreur sera précisée.

76 Enfin, quelques erreurs, relativement peu nombreuses, portent sur la suppression d'une lettre à l'intérieur d'un graphème complexe (composé de plusieurs lettres). Lorsque cette suppression entraine une modification de la valeur sonore du graphème, il semble qu'il soit plus pertinent d'analyser cette erreur au niveau de la lettre et non comme une substitution d'un graphème par un autre. L'annotation est donc portée par la lettre.

77 Comme nous l'avons évoqué précédemment, la détection d'erreur et la recherche de la forme attendue se font au niveau de la forme, une annotation des formes doit donc également être faite. Elles seront annotées selon leur écart à la norme. Elles peuvent ainsi être :

> normés, s'ils respectent la norme orthographique ;
> erronés de type orthographique, si la phonologie du mot transcrit est respectée mais pas la norme orthographique ;
> erronés de type phonographique, Si elles ne respectent pas le code phonographique, c'est-à-dire qu'elles ne transcrivent pas la forme phonologique du mot.

78 De plus, l'outil TreeTagger permettra de faire une annotation morphosyntaxique des formes précisant la catégorie et le lemme. Précisons tout de même que cette étape nécessitera à nouveau un travail plus approfondi étant donné le taux d'erreur de TreeTagger sur ce type de données.

79 Enfin, la complexité, la multiplicité et le caractère non obligatoire des signes de ponctuation font que la modélisation des erreurs portant sur ces signes n'est pas encore parvenue à son terme. La place attribuée aux signes graphiques et à l'annotation des erreurs les concernant doit encore être réfléchie.

3.5.2 Élaboration du programme

80 Une fois le modèle d'annotation élaboré, il nous est alors possible de concevoir un outil d'annotation automatique des erreurs d'orthographe. Comme annoncé précédemment, nous ne présenterons ici que le module d'annotation des formes erronées de type orthographique.

81 Possédant la forme en corpus et la forme attendue, il est alors possible de les comparer pour détecter et annoter les erreurs. Une segmentation en graphèmes de chacune de ces formes est d'abord nécessaire. Nous nous inspirons ici des règles développées pour le système LIA_PHON. Les règles conçues pour ce système sont des règles contextuelles, respectant la prononciation du français. Chaque règle permet d'identifier un graphème et sa réalité sonore selon le contexte dans lequel il se place.

82 Lors de cette étape, sont calculées pour chaque graphème la valeur du phonème qu'il transcrit, sa nature (voyelle, consonne, etc.), mais également sa fonction (lexicale ou grammaticale, selon que le graphème se rapporte au radical de l'unité lexicale ou à ses flexions). La valeur de cette dernière est calculée de manière automatique en calculant le radical à partir de la catégorie et du lemme donnés par TreeTagger.

83 Après repérage des formes comportant des erreurs (étape 1) et recherche de la forme normée attendue (étape 2), les deux formes peuvent être comparées et annotées graphème par graphème (étape 3).

84 Dans les cas comme *cha* (forme normée : *chat*) où le graphème *t* a été omis par l'enfant, ce graphème sera mentionné comme omis (avant-dernière ligne du tableau 5).

Tableau 5. Annotation de la forme *cha*

Étape 1	**Forme erronée**			
	cha			
Étape 2	**Forme attendue**		**Lemme**	**Catégorie**
	chat		chat	NOM
Étape 3	**En corpus**	**Attendu**	**Annotation**	
	cha	chat		
	ch	ch	<grapheme ecart="“**normé**” fonction="“lexicale”" sonorite="“consonne”" phono="“ch”">ch</grapheme>	
	a	a	<grapheme ecart="“**normé**” fonction="“lexicale”" sonorite="“voyelle”" phono="“aa”">a</grapheme>	
		t	<grapheme ecart="“**omis**” fonction="“lexicale”" sonorite="“muet”" valeur="“t”"></grapheme>	

85 Si un graphème supplémentaire a été ajouté, comme dans l'exemple *soire* (forme normée : *soir*), il sera annoté comme inséré (avant-dernière ligne du tableau 6).

Tableau 6. Annotation de la forme *soire*

Étape 1	**Forme erronée**
	soire

Étape 2	**Forme attendue**		**Lemme**	**Catégorie**
	soir		soir	NOM
Étape 3	**En corpus**	**Attendu**	**Annotation**	
	soire	soir		
	s	s	<grapheme ecart="**normé**" fonction="lexicale" sonorite="consonne" phono="ss">s</grapheme>	
	oi	oi	<grapheme ecart="**normé**" fonction="lexicale" sonorite="voyelle" phono="wwaa">oi</grapheme>	
	r	r	<grapheme ecart="**normé**" fonction="lexicale" sonorite="consonne" phono="rr">r</grapheme>	
	e		<grapheme ecart="**inséré**" fonction="lexicale" sonorite="muet">e</-grapheme> </grapheme>	

86 Enfin, lorsqu'un graphème est substitué à un autre gra-phème, mais que la valeur phonique est respectée, comme dans *tonba* (forme normée : *tomba*), il sera annoté comme substitué de type orthographique (huitième ligne du tableau 7).

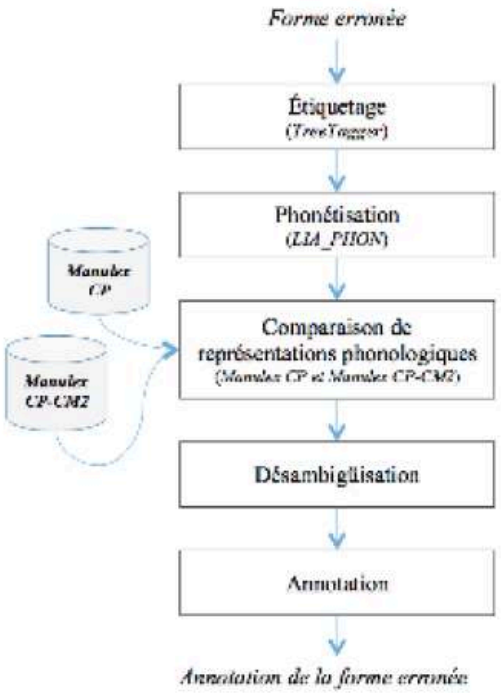
Tableau 7. Annotation de la forme *tonba*

Étape 1	**Forme erronée**			
	tonba			
Étape 2	**Forme attendue**		**Lemme**	**Catégorie**
	tomba		tomba	VERBE
Étape 3	**En corpus**	**Attendu**	**Annotation**	
	tonba	tomba		
	t	t	<grapheme ecart="**normé**" fonction="lexicale" sonorite="consonne" phono="tt">t</grapheme>	

	on	om	<grapheme ecart="**orthographique**" fonction="lexicale" sonorite="voyelle" phono="on" valeur="om">on</grapheme>
	b	b	<grapheme ecart="**normé**" fonction="lexicale" sonorite="consonne" phono="bb">b</grapheme>
	a	a	<grapheme ecart="**normé**" fonction="grammaticale" sonorite="voyelle" phono="aa">a</grapheme>

87 Le programme élaboré (figure 3) prend ainsi une forme en entrée et produit son annotation graphème par graphème en sortie. La forme est d'abord étiquetée à l'aide de TreeTagger, si elle est étiquetée <unknown> par cet outil, alors une représentation phonologique de cette forme est produite par LIA_PHON. Puis, celle-ci est comparée aux représentations phonologiques des formes de la liste Manulex CP. Si aucun correspondant n'est trouvé, on procède de même avec la liste Manulex CP-CM. Dans le cas contraire, après désambiguïsation de la forme attendue, la forme erronée est comparée graphème par graphème à celle-ci et une annotation est produite. Cette annotation spécifie pour chaque graphème, son écart à la norme, la valeur du graphème (forme en corpus), la valeur attendue (forme normée attendue), la valeur phonique, la sonorité et la fonction.

Figure 3. Processus d'annotation d'une forme erronée



4. Perspectives

88 Cette étude constitue une première approche de l'usage de méthodes issues du TAL sur des textes scolaires très peu normés. Elle semble confirmer l'hypothèse que la connaissance du contexte de production permet d'améliorer sensiblement la qualité du processus d'annotation par le recours de ressources spécifiques au contexte. Bien entendu, une évaluation du dispositif sur un échantillon plus large reste à mener.

89 Dans notre perspective d'aide à l'annotation et à l'exploitation du corpus, une autre piste autour de l'apport du TAL est à l'étude. Elle consiste, lors de l'étape de transcription, à proposer manuellement un « corrigé » de la production. Ce corrigé sera par la suite aligné automatiquement avec la production. L'annotation automatique s'appuiera alors sur des comparaisons entre production de l'élève et correction.

90 Un autre aspect important de notre recherche concerne les possibilités d'exploitation de ce corpus par les chercheurs et les enseignants. Couplée à un module de requêtes, l'annotation va permettre de procéder à différentes recherches à travers le corpus. Il sera ainsi possible de rechercher l'ensemble des graphies employées pour transcrire un même phonème par les scripteurs débutants, leurs fréquences, le taux d'erreurs lié à ce phonème ou encore les contextes favorisant l'échec ou la réussite de l'écriture d'une graphie. Là également, nous comptons nous appuyer sur le « corrigé » pour étendre les capacités d'exploration du corpus. Ainsi, à l'instar du système Exxelant (Antoniadis *et al.*, 2007), les requêtes pourront porter aussi bien sur la production de surface que sur la proposition de corrigé.

91 Cependant, un outil d'exploration n'a de sens que s'il répond aux attentes de ses utilisateurs. C'est pourquoi, une collaboration avec des chercheurs et des enseignants va débuter prochainement. Une mise à disposition du corpus de CP accompagné d'une première version de l'outil exploratoire sera proposée aux enseignants et chercheurs impliqués dans le projet début 2016. Le corpus sera complété au fur et à mesure de la recherche et les retours sur l'outil serviront de base à ses évolutions.

BIBLIOGRAPHIE

Ågren M. (2008). *À la recherche de la morphologie silencieuse. Sur le développement du pluriel en français L2 écrit*, Études Romanes de Lund 84, Thèse de doctorat, Université de Lund.

Andersen H. L., Leblay C., & Auriac-Slusarczyk E. (2010). « Pourquoi travailler sur un corpus commun ? Pourquoi travailler de manière pluridisciplinaire ? », *Synergies Pays Scandinaves* 5 : 17-30.

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 Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, Grenoble, France, 95-108.

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, 770-779.

Béchet F. (2001). « LIA_PHON - Un système complet de phonétisation de textes », *Traitement Automatique des Langues (T.A.L.)* 42(1) : 47-67.

Blanche-Benveniste C. & Chervel A. (1969, éd. augmentée, 1978). *L'orthographe.* Paris : Maspero.

Bouysse V. (2006). « L'école maternelle, un modèle quasi mythique », *La revue de l'inspection générale* 3 : 18-25.

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

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

Carlson A., & Fette I. (2007). « Memory-based context-sensitive spelling correction at web scale », Sixth *International Conference on Machine Learning and Applications, 2007. ICMLA 2007.* IEEE, 166-171.

Chanier T., (1992). « Perspectives de l'apport de l'EIAO dans l'apprentissage des langues étrangères : modélisation de l'apprenant et diagnostic d'erreurs », *Revue de liaison de la recherche en Informatique Cognitive des Organisations* 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.

Damerau F. J. (1964). « A technique for computer detection and correction of spelling errors ». *Communications of the ACM 7*(3) : 171-176.

Elalouf M.-L. (dir.) (2005). Écrire entre 10 et 14 ans. *Un corpus, des analyses, des repères pour la formation*, SCérén, CRDP de Versailles.

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

Goigoux R. (2016). *Étude de l'influence des pratiques d'enseignement de la lecture et de l'écriture sur la qualité des premiers apprentissages*, Rapport de recherche, Ifé, http://ife.ens-lyon.fr/ife/recherche/lire-ecrire/rapport/rapport.

Granger S. (2009). « The contribution of learner corpora to second language acquisition and foreign language teaching », *Corpora and language teaching* 33 : 13.

Fayol M., & Jaffré J. P. (2008). *Orthographier*. Paris, PUF.

Kernighan M. D., Church K. W., & Gale W. A. (1990, August). « A spelling correction program based on a noisy channel model », *Proceedings of the 13th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, 205-210.

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. Url : http://www.u-grenoble3.fr/ponton/perso/docs/TALN07.pdf.

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

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(1) : 156-166.

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

New B., Pallier C., Ferrand L. & Matos R. (2001). « Une base de données lexicales du français contemporain sur internet : LEXIQUE™//A lexical database for contemporary french : LEXIQUE™ », *L'année Psychologique* 101(3) : 447-462. http://www.lexique.org.

Ortéga É., & Lété B. (2010). « eManulex : Electronic version of Manulex and Manulex-infra databases », http://www.manulex.org.

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.

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

Soulé Y., Kervyn B., Geoffre T., Chabanne J.C. (2015). « Évaluer la production d'écrit en fin du cours préparatoire (première primaire). De l'élaboration d'une épreuve de test à l'analyse des résultats obtenus », in Falardeau E., Lefrançois P., Dumortier J.L., Dolz J. (dir.). *L'évaluation en classe de français, outil didactique et politique*. AIRDF. Namur : Presses Universitaires de Namur.

Smith N., & McEnery T. (1998). « Issues in Transcribing a Corpus of Children's Hanwritten Projects », *Literary and linguistic computing*, 13(4) : 217-225.

Walter H. (1976). *La dynamique des phonèmes dans le lexique français contemporain*. Librairie Droz.

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.

NOTES

1. http://www.lancaster.ac.uk/fass/projects/lever/index.htm.
2. Pour plus de précision sur cette méthode se référer à l'article de Soulé Y., Kervyn B., Geoffre T., Chabanne J.C. (2015).
3. Ce chiffre est différent de celui avancé précédemment car toutes les productions n'ont pas encore pu être rassemblées.
4. Pour simplifier les exemples, dans la suite de cet article nous nous baserons sur une version des productions exemptes de ces balises.
5. À noter que les passages complètement illisibles sont repérés avec une simple balise <illisible/>.
6. Antidote Prisme (version 5) [logiciel]. Montréal, Canada : Druide informatique.
7. Cordial 2010 (versions professionnel et standard) [logiciel]. Toulouse, France : Synapse Développement. http://www.synapse-fr.com.

8. Pour le moment, les fréquences d'hypersegmentation et d'hyposegmentation n'ont pas encore été mesurées sur l'ensemble du corpus, mais on peut déjà observer une tendance plus importante à l'hyposegmentation qui reste à vérifier.

9. Pour traiter les erreurs commises par les élèves de CP, nous nous concentrons sur les aspects phonographiques, les notions de morphogrammes et de morphographie ne sont donc pas employées dans cet article.

10. Données valables au moment de l'élaboration de cette liste.

11. Parmi les quelques 138.582 formes que compte cette ressource, seules 19.328 formes sont contenues dans la liste Manulex CP et 76.676 dans la liste Manulex CP-CM2, soit des taux de recouvrement respectivement de 13,9 % et de 55,3 %. Précisons également que les listes MAnulex contiennent quelques formes exclues de la ressource Lexique 3 ; ce sont principalement des noms propres et les formes composées par ajout de tirets ou d'apostrophes, non inclues dans cette ressource.

12. Cette notion est empruntée à C. Blanche-Benveniste et A. Chervel (1969) après adaptation. Il s'agit d'un graphème n'ayant pas de valeur phonique mais ayant une influence sur la valeur phonique des graphèmes voisins.

RÉSUMÉS

Le travail présenté dans cet article s'inscrit dans une recherche qui a pour but la constitution d'un corpus scolaire et le développement d'un outil d'aide à son exploitation à partir de l'annotation de phénomènes linguistiques saillants. Nous nous concentrerons ici sur les écrits produits en fin de classe de CP par des scripteurs encore débutants. L'objet de ce travail est d'explorer les possibilités qu'offre le traitement automatique des langues pour appréhender ces écrits particulièrement éloignés de la norme. L'hypothèse est que la connaissance du contexte de production facilite ce processus. Nous mesurons cet apport au travers d'un exemple de traitement, à savoir le développement d'un outil d'aide à l'annotation de certaines erreurs orthographiques. Après une rapide présentation du projet et des caractéristiques du corpus élaboré, l'article propose un exposé détaillé du module d'annotation de ces erreurs. Il en expose la méthode d'identification et de correction au moyen d'une ressource lexicale de formes phonologiques ainsi que le modèle d'annotation élaboré.

Constituting a school corpora with NLP

Our study takes part in a project which aims at elaborating a large corpus of school texts and at developing a linguistic tool facilitating its exploitation. In this article, the focus is put on texts written by novice writers: children at the end of the first year of schooling (6-7 year-old). This study explores possibilities given by natural language processing to annotate non-normed school corpora. Our hypothesis is that the knowledge of the context can ease this process. We measure this contribution through an example of processing, the development of a help tool for specific spell checking. First the project and specificities of the corpus are presented; then, the spell errors annotation module is detailed, both the spell checking methods on the basis of a phonological lexical resource and the annotation model.

INDEX

Mots-clés : traitement automatique des langues, corpus de textes scolaires, orthographe
Keywords : natural language processing, school corpora, spell checking

AUTEURS

CLAIRE WOLFARTH

Université Grenoble Alpes

CLAUDE PONTON

Université Grenoble Alpes

CORINNE TOTEREAU

Université Grenoble Alpes