Évolution de la gestion de l'orthographe en production de textes aux cycles 2 et 3 : apports du TAL

Claude Ponton, Université Grenoble Alpes, Lidilem
Catherine Brissaud, Université Grenoble Alpes, Lidilem
Claire Wolfarth, Université Grenoble Alpes, Lidilem

Résumé. Nous présenterons une analyse assistée par des outils informatiques de données recueillies dans le cadre du projet E-Calm, pour les niveaux allant du CP en 6e. Nous étudierons l'évolution de la segmentation en mots, du traitement du verbe et de la morphologie de l'adjectif dans des productions narratives. Des pistes didactiques seront proposées.

Abstract. We will present a computer-assisted analysis of data collected in the framework of the E-Calm project, for levels ranging from grade 1 to grade 6. We will study the evolution of word segmentation, verb processing and adjective morphology in narrative productions. Didactic approaches will be presented.

Mots clés. corpus scolaire longitudinal ; évolution de la gestion de l'orthographe ; productions d'écrit ; traitement automatique des langues

L'un des objectifs du projet ANR E-Calm[3] est la caractérisation des écrits d'élèves et d'étudiants du point de vue de l'orthographe. Dans cet article, nous nous intéressons plus particulièrement à la gestion de l'orthographe en production d'écrit pour les cycles 2 et 3 qui reste encore peu documentée (Elalouf, 2005).

Les données de l'étude

Le corpus complet sur lequel servant de base à cette étude est constitué d'écrits d'élèves du CP à la 6e répartis de la manière suivante :

Niveau	Nombre de textes	Nombre de mots	Longueur moyenne des textes
CP	373	9 965	26,72
CE1	373	24 826	66,56
CE2	373	43 373	116,28
CM1	373	52 601	141,02
CM2	337	54 484	173,54
6e	161	28 893	274,74
Total	**1990**	**218 142**	**109,62**

Le sous-corpus CP-CM2 correspond au corpus Scoledit[4] (Wolfarth *et al.*, 2018a) qui comporte des textes narratifs produits par les mêmes enfants suivis du CP au CM2 (corpus

[3] http://e-calm.huma-num.fr/le-projet/
[4] http://scoledit.org/scoledition/

longitudinal). Le sous-corpus de 6e est composé d'écrits de collégiens recueillis par les équipes Clesthia dans le projet Ecriscol[5] (Doquet *et al*., 2017) et CLLE[6] (Garcia-Debanc *et al*., 2018).

Précisons ici que l'étude ci-après portant sur la segmentation est antérieure et ne porte que sur 333 textes du CP ; une mise à jour sera menée ultérieurement.

Pour assister la description linguistique de ce corpus, une chaine de traitement semi-automatique (Wolfarth *et al*., 2018b) a été développée dans le cadre du projet E-Calm. Cette chaine prend en entrée les transcriptions et une version normée des textes (Ho-Dac *et al*., 2020). Des comparaisons entre mot transcrit et mot normé sont ensuite opérées à différents niveaux : graphique, phonologique, lexical, morphosyntaxique… Ces comparaisons fournissent ainsi des éléments de description sur les réussites et les difficultés des élèves comme, par exemple, sur la segmentation en mots, la morphologie verbale ou la morphologie adjectivale.

La segmentation en mots au CP

Cette première exploitation du corpus se concentre sur l'analyse des segmentations non conventionnelles. En effet, connaitre les correspondances entre graphème et phonème (principe phonographique) ne suffit pas pour écrire le français ; il y a aussi le principe sémiographique qui renvoie au sens et qui inclut la délimitation des mots. Cela implique de se construire une conception de ce qu'est un mot quand on écrit. Ainsi la séparation entre les mots, représentées par des blancs ou espaces, n'est pas naturelle, notamment parce qu'elle ne correspond pas à la segmentation de la parole. Il faudra plusieurs années à l'élève de l'école élémentaire pour en venir à bout, notamment en production d'écrit. Cela dit, la phrase dictée au début et en fin de CP *Tom joue avec le rat* dans le cadre de l'étude *Lire et écrire au CP* a permis de mesurer les progrès des élèves sur la seule année de CP : « En septembre, 15,2 % des élèves n'ont pas écrit la phrase et 61 % l'écrivent en un seul bloc. 23,8 % des élèves commencent à segmenter la phrase en mots (de 2 à 6 mots). En juin, près de 90 % des élèves segmentent la phrase en mots (4, 5 ou 6 segments) » (Goigoux, 2016, p. 170).

La segmentation dans l'écriture est représentée visuellement par l'insertion d'espaces vides (ou de blancs) entre les mots. L'hyposegmentation consiste à laisser moins d'espaces vides qu'attendu, c'est-à-dire à agglutiner les mots plus que nécessaire (par exemples *senva* pour *s'en va*) ; inversement, l'hypersegmentation consiste à insérer des blancs à l'intérieur des mots et donc à fragmenter les mots graphiques (par exemple *a ve* pour *avait*).

Les deux phénomènes peuvent se combiner, par exemple *par seque* pour *parce que*.

Présence de l'hypo et de l'hypersegmentation dans le corpus Scoledit

Un peu plus d'un tiers des 333 textes de CP analysés ne présente aucun problème de segmentation (37,54%). L'hyposegmentation est le phénomène le plus fréquent dans notre corpus, conformément aux recherches antérieures (Ferreiro et Pontecorvo, 1996 ; Piacente et Querejeta, 2012 ; Ugarte et Argüero, 2017). On en trouve au moins un exemple dans 54,35% des 333 textes. Les textes présentant au moins un cas d'hypersegmentation sont moins nombreux : ils représentent 23,42% du corpus.

Enfin, 8,11% des textes présentent au moins un cas mêlant hypo et hypersegmentation.

[5] http://syled.univ-paris3.fr/ecriscol/CORPUS-TEST/
[6] http://redac.univ-tlse2.fr/corpus/resolco.html

Nombre d'éléments impliqués

S'agissant des cas d'hyposegmentation, les cas composés de deux éléments ou mots graphiques réunis en un seul mot prédominent (83,69%). Les cas d'hyposegmentation impliquant plus de deux éléments représentent 16,06% des occurrences.

En ce qui concerne les cas d'hypersegmentation, on constate une nette prévalence d'une hypersegmentation en deux éléments. Ce cas représente 98,15% des hypersegmentations dans le corpus.

Catégories grammaticales concernées par les phénomènes d'hyposegmentation

Les éléments de type fonctionnel/grammatical (articles, pronoms, prépositions et conjonctions) ont été distingués des éléments de type lexical (noms, adjectifs, verbes et adverbes). Le cas le plus fréquent d'hyposegmentation est composé d'un élément fonctionnel et d'un élément lexical (83,95%) ; la configuration premier élément = pronom personnel et second = verbe prédomine (30%), par exemple *la* (l'a), *sest* (s'est), *cerévei* (se réveillent), *seléve* (se lève), *onva* (on va). 7% des cas consistent en une préposition suivie d'un nom, par exemple *partere* (par terre).

L'hypersegmentation : un premier élément de type fonctionnel

Le premier élément est le plus souvent de type fonctionnel : 10,18% des cas contiennent « a » comme premier élément (*a ve* (avait), *a pre* (après)) ; 8,33% des cas présentent « c' » comme premier segment (*c'est* (ses)) ; 6,48% des cas impliquant le segment « et » comme premier (*et te* (était), et 6,48% avec le segment « ton », par exemple, *ton be* (tombe).

Conclusion

La proportion de mots graphiques correctement séparés par des blancs est révélatrice de la conscience qu'a l'élève de l'organisation spatiale de l'écrit et des problèmes que pose le passage de la chaine orale à la chaine écrite. Cette observation a déjà été réalisée dans d'autres langues romanes. Ainsi, ces données semblent confirmer que les segments que les élèves ont tendance à accoler plus qu'ils ne le devraient ou, inversement, à segmenter à l'excès, ne sont pas aléatoires.

Le traitement du verbe

Le corpus de la présente étude contient 42 807 formes verbales (hors participes) représentant 17,95% des mots. Cette proportion est relativement stable sur l'ensemble des niveaux scolaires. Du fait, en partie des consignes et du genre narratif des textes, les élèves mobilisent principalement 4 temps ou tiroirs verbaux (présent et imparfait de l'indicatif, infinitif présent, passé simple) qui représentent à eux seuls 97,23% de l'ensemble. On note également une progression régulière dans la réussite de ces formes verbales puisque l'on passe de 25,94% de formes correctement orthographiées au CP à 68,21% en 6e. Lors de nos études précédentes (Wolfarth *et al*., 2017b ; Lavieu-Gwozdz *et al.*, à paraitre) consolidées par l'étude actuelle, deux constats principaux s'imposent.

Si des progrès sont constatés sur les quatre tiroirs verbaux principaux, on note que l'imparfait et surtout le passé simple sont ceux pour lesquels ces progrès sont les plus lents.

Si la gestion des bases lexicales s'améliore rapidement (de 54,51% d'erreurs au CP à 11,59% en 6e), la gestion des désinences verbales reste une difficulté persistante puisqu'elle constitue encore 23,98% d'erreurs en 6e.

Le traitement de l'adjectif

Avec un total de 9 113 formes, les adjectifs représentent 3,82% des formes de notre corpus alors qu'à titre de comparaison, les adjectifs représentent 6,96% des formes de Manulex CP-CM2[7]. Cette catégorie grammaticale semble donc peu mobilisée par les élèves et ce, quel que soit le niveau : de 3,49% en 6e à 4,54% au CP. Il est à noter que la consigne du CP demandant aux élèves de raconter l'histoire d'un petit chat implique un suremploi de l'adjectif « petit » (près de 76% des adjectifs mobilisés) faussant ainsi l'étude comparative. De ce fait, nous ne tiendrons pas compte du niveau CP dans cette étude.

La réussite orthographique s'améliore d'un niveau à l'autre passant de 52,99% de réussite en CE1 à 74,13% en 6e. On note toutefois une stagnation des progrès entre le CM2 (76,42% de réussite) et la 6e (74,13%). Il est difficile d'expliquer cette stagnation sans une étude plus précise. Toutefois, on peut noter que les textes de CE1-CM2 sont produits à partir d'une même consigne sur la même cohorte d'élèves alors que les textes de 6e sont issus de consignes et d'élèves différents.

À l'aide de programmes de traitement automatique des langues, nous avons opéré un découpage des formes adjectivales en base, flexion de genre et flexion de nombre. Ce découpage nous a permis de distinguer les erreurs portant sur la base de celles portant sur les différentes flexions.

Si les réussites sur les bases lexicales s'améliorent progressivement, les erreurs restent plus élevées que sur les flexions, contrairement à ce qui est observé pour les verbes. En 6e, par exemple, 13,80% des adjectifs présentent encore une erreur de type lexical.

Concernant les accords en genre ou en nombre, on remarque une réussite supérieure sur les valeurs par défaut (masculin, singulier) qui ne nécessitent pas de marques flexionnelles. Tous niveaux confondus, le pluriel est la marque qui pose le plus de difficulté aux élèves puisque pour les adjectifs attendus au pluriel, 6 sur 10 sont erronés. Il conviendra toutefois dans une étude ultérieure d'affiner la description de ces erreurs en distinguant par exemple celles où l'élève produit des formes existantes mais non accordées (ex. *gentil* au lieu de *gentils*) de celles pour laquelle la forme produite est inexistante (ex. *nouvèl* au lieu de *nouvelle*).

Perspectives didactiques

Le corpus présenté permet la mise en évidence d'une part des progrès continus des élèves de l'école élémentaire en orthographe en production d'écrit, d'autre part des zones du système orthographique qui restent longtemps en construction, tel le marquage morphographique des catégories du verbe et de l'adjectif.

Il permet aussi de distinguer les phénomènes récurrents à un niveau donné de ceux qui le sont moins et de repérer les élèves les moins avancés, par exemple en observant la segmentation en mots ou les choix graphiques concernant l'imparfait.

Ainsi, les résultats obtenus permettent de mieux cibler les tâches proposées aux élèves et d'effectuer des choix sinon rentables du moins raisonnables. Par exemple, concernant la segmentation, les suites déterminant + nom et pronom + verbe seraient à travailler en priorité, si possible en lien avec les activités grammaticales.

Références bibliographiques

Doquet, C., David, J. et Fleury, S. (dir.) (2017). Spécificités et contraintes des grands corpus de textes scolaires : problèmes de transcription, d'annotation et de traitement. *Corpus, 16*.

[7] http://www.manulex.org/fr

Doquet, C., Enoiu, V., Fleury, S. et Maziotti, S. (2017). Problèmes posés par la transcription et l'annotation d'écrits d'élèves. *Corpus*, *16*.
Elalouf, M.-L. (2005). *Écrire entre 10 et 14 ans un corpus, des analyses, des repères pour la formation*. Paris : Canopé - CRDP de Versailles.
Elalouf, M.-L. (2011). Constitution de corpus scolaires et universitaires : vers un changement d'échelle ? *Pratiques*, *149-150*, 56-70.
Garcia-Debanc, C., Ho-Dac, L.-M., Bras, M. et Rebeyrolle, J. (2018). Vers l'annotation discursive de textes d'élèves. *Corpus*, *16*.
Goigoux, R. (éd.) (2016), *Étude de l'influence des pratiques d'enseignement de la lecture et de l'écriture sur la qualité des premiers apprentissages.* Institut Français de l'Éducation, [consulté le 19/03/2021, http://ife.ens-lyon.fr/ife/recherche/lire-ecrire/rapport/rapport-lire-et-ecrire].
Ho-Dac, M., Fleury, S. et Ponton, C. (2020). É:Calm Resource: a Resource for Studying Texts Produced by French Pupils and Students. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. 4327–4332. Marseille. 11-16 May 2020.
Lavieu-Gwozdz, B., Vinel, E., Goossens, V. et Brissaud, C. (à paraitre). Cartographie des usages et des erreurs orthographiques sur les verbes dans des récits écrits par des élèves de 6 ans à 15 ans. *Langue Française*. « Écrire de l'école à l'université : corpus, traitements, analyses outillées ».
Ponton, C., Gutiérrez-Cáceres, R., Teruggi, L., Farina, E., Brissaud, C. et Wolfarth, C. (à paraitre). Scolinter : un corpus trilingue. L'exemple de la segmentation en mots. *Langue Française*. « Écrire de l'école à l'université : corpus, traitements, analyses outillées ».
Wolfarth, C., Brissaud, C. et Ponton, C. (2018a). Transcrire et normer un corpus scolaire : pour quelles analyses ? *Dyptique*, *36*, Presses Universitaires de Namur, 121-145
Wolfarth, C., Ponton, C. et Brissaud, C. (2018b). Gestion de la morphographie verbale en production d'écrits : que peut nous apprendre un corpus longitudinal ? *Repères*, *57*, 209-226.