Du TAL dans les écrits scolaires : premières approches

Claire Wolfarth[1] Claude Ponton[1] Catherine Brissaud[1]
(1) Lidilem, CS40700 - 38058 Grenoble cedex 9, France
Claire.Wolfarth@univ-grenoble-alpes.fr, Claude.Ponton@univ-grenoble-alpes.fr, Catherine.Brissaud@univ-grenoble-alpes.fr

RESUME

Dans cet article est présentée une première approche de l'usage de méthodes issues du TAL pour exploiter des textes scolaires, très peu normés. Il permettra d'envisager la spécificité de ces écrits à travers la présentation du corpus étudié avant de se pencher sur les premières hypothèses de traitement automatique en vue d'une annotation des erreurs qui le composent. Y seront également exposés les objectifs de ce travail et la portée attendue.

ABSTRACT

Some NLP in school texts corpora : first hypothesis
In this article, a first approach of non-normed school corpora treatment by using methods from NLP will be exposed. Specific features of this type of texts will be shown by presenting the corpus and first hypothesis to achieve its annotation in term of errors. Main goals of this work will also be presented.

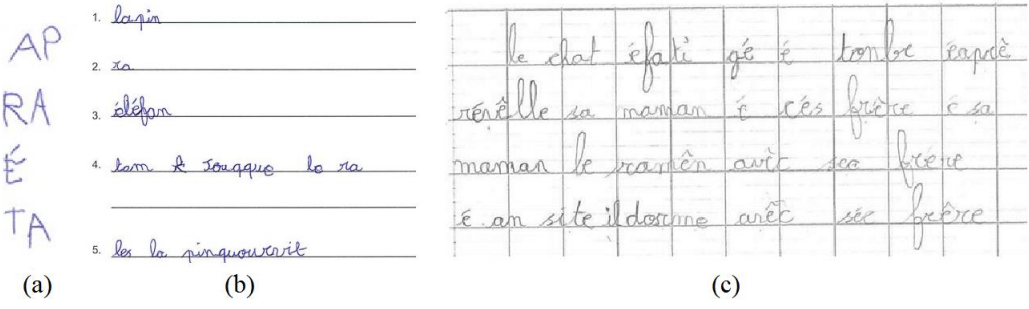
MOTS-CLES : corpus scolaires, annotation automatique, erreurs d'apprenants.
KEYWORDS: school texts corpora, automatic annotation, learner errors

1 Introduction

Le travail proposé s'inscrit dans un projet visant à élaborer un large corpus d'écrits d'apprenants accompagné de modules spécifiques d'analyse afin d'outiller la didactique de l'écriture. Ce corpus devrait permettre la description des structures linguistiques utilisées par des élèves en cours de construction de leurs apprentissages de l'écrit à différents niveaux de fonctionnement linguistique (morphographie, syntaxe, lexique, structuration du discours) dans le but de rendre compte des évolutions des procédés d'écriture à différents moments de la scolarisation à l'école primaire. Si certains travaux se sont déjà penchés sur ces questions de description (Elalouf, 2005 ; Auriac-Slusarczyk, Gunnarsson, 2014), ils s'appuient généralement sur des corpus restreints. A terme, la constitution d'un tel corpus et son exploitation devraient permettre de déboucher sur le développement d'activités didactiques adaptées.

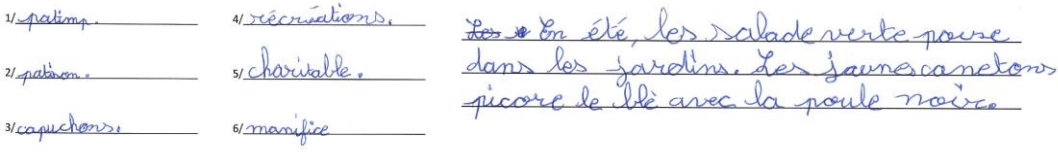
Le corpus visé rassemblera différents textes et dictées produits selon un même protocole par les mêmes élèves à des niveaux d'apprentissage différents. Ces productions sont recueillies à chaque fin d'année scolaire du CP au CM2 (2014-2018) pour les textes, en début d'année de CP et en fin d'année de CP et de CE1 pour les dictées. Les élèves concernés sont répartis dans 55 écoles de 5 académies (Grenoble, Clermont-Ferrand, Bordeaux, Lyon et Toulouse soit 1.230 élèves de CP). À terme, ce corpus devrait contenir plus de 3000 productions, ce qui représente, pour le domaine des

corpus scolaires, un corpus longitudinal de grande taille. Actuellement, seuls les recueils en classe de CP et de CE1 ont été réalisés et seules celles de CP ont été transcrites. Ci-dessous, différents exemples de productions sont donnés.

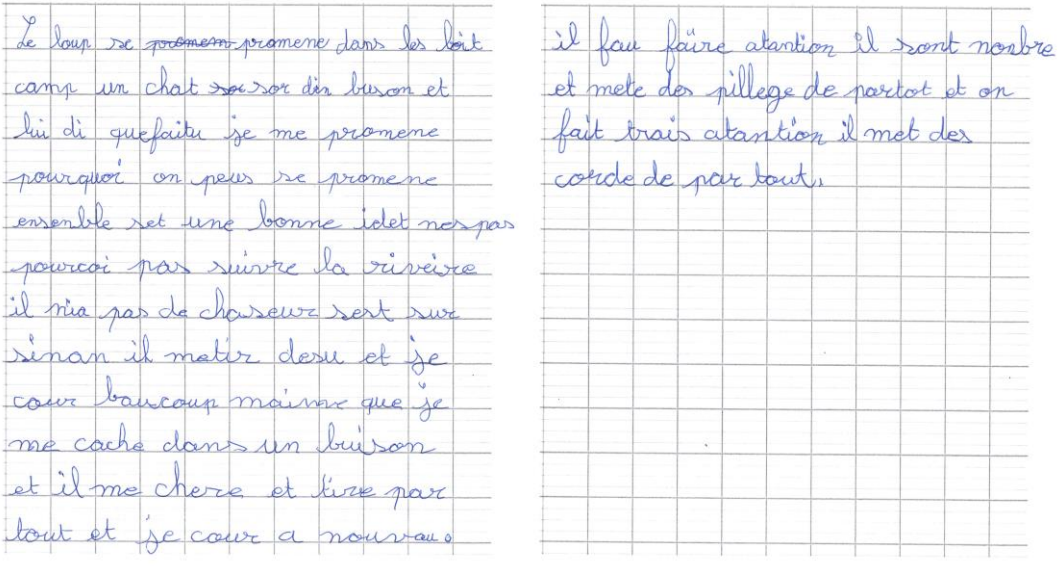


(a) (b) (c)

EXEMPLE 1 : Productions de l'élève 1558 en classe de CP. (a) Dictée produite au mois de septembre. Les mots et phrases dictés sont "lapin", "rat", "éléphant" et "Tom joue avec le rat". (b) Dictée produite au mois de juin. Les mots et phrases dictés sont "lapin", "rat", "éléphant", "Tom joue avec le rat" et "les lapins courent vite". (c) Texte produit au mois de juin. L'enfant devait écrire un texte à partir de 4 images.



EXEMPLE 2 : Dictée produite par l'élève 1558 au mois de juin de la classe de CE1. Les mots et phrases dictés sont "patin", "pâtisson", "capuchon", "récréation", "charitable", "magnifique" et "En été, les salades vertes poussent dans les jardins. Les jeunes canetons picorent le blé avec la poule noire".



EXEMPLE 4 : Texte produit par l'élève 1558 au mois de juin de la classe de CE1. L'enfant devait écrire un texte à partir d'un personnage choisi parmi 4 vignettes.

Afin, de permettre une exploitation riche de ce corpus par les enseignants, les linguistes et les didacticiens, une annotation en terme d'erreurs et de phénomènes notables est prévue. Au vu de la taille du corpus, un recours à des méthodes et outils issus du traitement automatique des langues (TAL) est envisagée. Ces méthodes devraient contribuer à la fois à l'annotation et à l'exploitation du corpus. L'enjeu du projet global est donc triple : 1/ un enjeu linguistique de constitution d'une ressource outillée pour la recherche en linguistique ; 2/ un enjeu pour le TAL, de caractérisation et de modélisation de types d'écrits souvent très éloignés de la norme ; 3/ un enjeu pédagogique et didactique appuyé par la connaissance fine des acquis et difficultés, accessibles au travers d'un outil d'interrogation du corpus. Cet article présente les premières approches exploratoires effectuées dans cette optique sur les productions de CP. Elles constituent le point de départ de différents travaux en cours ou à venir aussi bien en synchronie sur les différents niveaux qu'en diachronie du CP au CM2.

2 État de l'art

Depuis les années 1980, le TAL est étroitement associé à la linguistique de corpus (Habert, Nazarenko, Salem, 1997 ; Kennedy, 1998) par les méthodes et outils qu'il offre pour concevoir et exploiter de grandes masses de données. On trouve dans la littérature autour du TAL un ensemble de travaux connexes au traitement des écrits scolaires, parmi lesquels différentes recherches concernent l'apport du TAL au domaine de l'apprentissage des langues avec notamment les différentes approches du traitement de l'erreur décrites dans l'ouvrage de Heift, Schulze (2007). Pour le français, le projet Freetext (Granger, Vandeventer, Hamel, 2001) mène un travail autour de la

détection automatique d'erreurs basée sur le corpus d'apprenants FRIDA. Ce corpus est également au cœur du projet Exxelant (Antoniadis, Ponton, Zampa, 2010) avec le développement d'un système d'interrogation portant à la fois sur les productions et les corrections.

Même si elle s'adresse le plus souvent à des scripteurs experts, la correction automatique de textes est également un domaine important du TAL proposant des approches variées (Kukich, 1992) potentiellement intéressantes pour le traitement des écrits scolaires. Le traitement automatique d'écrits peu normés est un domaine plus récent en TAL avec notamment les travaux autour de corpus SMS ou issus des réseaux sociaux. Bien que spécifique, ce type de corpus présente des similarités avec les corpus scolaires puisque l'on y retrouve, entre autres, des problématiques liées à la proximité avec l'oral ou à la segmentation en mots (Fairon, Klein, Paumier, 2006). Toujours dans le domaine de ces écrits peu normés, signalons également le système d'étiquetage morphosyntaxique MElt (Denis, Sagot, 2012) développé spécifiquement pour ce type d'écrits et notamment appliqué à des textes provenant de forums en ligne (Baranes, 2012).

3 Spécificités du corpus, hypothèses et premières approches

Outre l'étape de transcription[1] non triviale de ce type d'écrits, son caractère très peu normé constitue un défi pour le TAL. Bien que de nombreux travaux se soient intéressés aux corpus peu normés, il n'est pas possible de transposer ces travaux à notre corpus sans les adapter. En effet, comme le montre les exemples de production précédents, celui-ci présente de nombreuses particularités qui soulèvent des problèmes spécifiques. Pour le moment, seules les productions de CP ont été entièrement transcrites mais nous n'avons pas encore décrit finement, et encore moins quantifié, les phénomènes rencontrés. Toutefois, parmi les erreurs spécifiques, deux grandes catégories semblent se dégager. La première concerne les erreurs de segmentation, témoins d'une connaissance encore floue des frontières de mots chez certains élèves. On donnera pour exemple la production de l'élève 1346 en classe de CP « Le petit chat sanva pan dan cesa maman dore. [...] » (*Le petit chat s'en va pendant que sa maman dort*.), où l'on notera les formes « sanva » et « cesa » dont les formes normées (interprétées) sont respectivement *s'en va* (3 formes) et *que sa* (2 formes) et les formes « pan dan » que l'on peut normaliser par l'unique forme *pendant*. Nous appellerons les premiers exemples des phénomènes d'hyposegmentation et le deuxième un phénomène d'hypersegmentation.

La deuxième catégorie concerne les erreurs de correspondance phonographique pour lesquels l'enfant transcrit le phonème attendu à l'aide d'une graphie non normée (e.g. « tonbe », attendu *tombe*) et d'omission de lettres muettes (e.g. « cha », attendu *chat*). Nos premières constatations, semblent laisser penser que, pour ces erreurs, alors que la norme orthographique n'est pas respectée, les informations phonographiques présentes associées au contexte fournissent des indices forts pour retrouver les formes correctes.

[1] Cette étape consiste en une saisie manuelle, et donc une interprétation, des productions dans une base de données. Plus de détails sur cette opération sont donnés dans (Wolfarth et al., 2016)

À ces deux catégories, il convient de rajouter un ensemble d'erreurs, assez présentes au CP, difficilement interprétables (e.g. élève 56 : « cisr un qui c unqior BouM Miaou q un »).

Au vu de ces spécificités et dans l'état actuel de nos connaissances, il ne nous semble pas envisageable de viser une quelconque automatisation de l'annotation. A l'instar de l'approche moins-disante de Kraif, Ponton (2007), nous situons donc dans une perspective d'aide à l'annotation et à l'exploitation. Nous comptons, d'une part, nous appuyer sur les traitements de plus bas niveau (souvent les plus robustes) pour, ensuite élargir progressivement le spectre de nos analyses. D'autre part, nous faisons l'hypothèse que la connaissance du contexte de production (niveau des scripteurs, consigne de rédaction…) permettra d'affiner la précision des analyses. Le test de cette hypothèse est la base du premier travail que nous avons effectué autour des erreurs à phonologie respectée et que nous présentons brièvement dans le paragraphe suivant.

3.1 Identification des formes normées par comparaison phonologique

En partant du constat qu'un nombre conséquent d'erreurs réalisées en classe de CP modifient la graphie du mot tout en conservant la majeure partie des informations phonologiques, une recherche des formes normées basée sur des comparaisons de formes phonologiques a été élaborée. Pour ce faire, chaque forme a été transcrite sous forme phonologique à l'aide de l'outil LIA_PHON (Béchet, 2001) puis comparée à différentes ressources lexicales plus ou moins proches du contexte de production), elles-mêmes converties sous forme phonologique à l'aide de LIA_PHON. Le résultat de cette comparaison est une liste de formes possibles qu'il reste à désambigüiser.

Parmi les différentes ressources lexicales testées, la ressource Manulex (Manulex CP et Manulex CP-CM2, Ortéga, Lété, 2010) e été retenue, ce choix fait suite à une première étude (Wolfarth et al., 2016) qui semble confirmer l'hypothèse que la connaissance du contexte de production permet d'améliorer sensiblement la qualité du processus d'annotation par le recours à des ressources spécifiques au contexte. En effet, contrairement aux lexiques généraux, le recours à des lexiques proches du vocabulaire utilisé en primaire permet une amélioration nette du processus de désambigüisation.

Cependant, pour un nombre certain de formes toutes les informations phonologiques ne sont pas disponibles, une telle méthode ne permet pas alors de retrouver la forme attendue. Au vu de ces limites et 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.

3.2 Annotation par alignement avec un « corrigé »

Si le recours au contexte de production améliore l'analyse, elle est donc loin d'être suffisante. L'idée que voudrions développer dans la suite de nos travaux est de s'appuyer sur une correction des productions pour pouvoir, à travers des comparaisons entre corrections et productions, proposer des analyses fines des erreurs. Si les corrections sont connues pour les activités de dictée, elles seront proposées manuellement lors de l'étape de transcription pour les productions. Cette analyse

comparative nécessitera au préalable une étape d'alignement entre correction et production. Nous comptons pour cela nous appuyer sur les marques phonologiques fortes présentes dans les textes. La comparaison entre les unités alignées se fera sur différents niveaux (graphiques, lexicaux, morphologiques, phonétiques, etc.) comme proposé dans le travail de Blanchard et al. (2009).

Un premier travail selon cette méthode est en cours sur les dictées réalisées au mois de juin en classe de CP et plus spécifiquement sur les deux phrases dictées ("Tom joue avec le rat" et "les lapins courent vite"). Dans ce cadre, un premier aligneur dictées/corrections ainsi qu'un premier module de comparaison sont en cours de développement. L'aligneur s'appuie sur les formes phonologiques et permet de produire des appariements entre segments comme [*courvit*] avec [*courent vite*]. La première version de l'analyseur de différence s'intéresse à la détection des cas d'hyper et d'hypo-segmentation, aux ajouts/omissions de mots, au niveau graphique (casse, différence de caractères) et au niveau phonologique (respect ou non de la phonologie). Un travail aux niveaux des graphèmes, des syllabes et de la morphologie est prévu.

4 Perspectives

Cette étude constitue une première approche de l'usage de méthodes issues du TAL sur des textes scolaires très peu normés. Les deux approches décrites précédemment sont en cours d'amélioration et d'évaluation. De ces différentes études sont attendus différents apports pour l'étude de l'apprentissage de l'écriture. En effet, il s'agira de mettre à disposition des communautés scientifiques et pédagogiques un corpus outillé, riche et unique pour sa taille et, surtout, pour son aspect longitudinal. D'autre part, le développement de méthodes et d'outils TAL spécifiques pour sa constitution et son interrogation permettra une approche fine de l'ensemble des données qui resterait difficile de manière manuelle.

Sur le plan linguistique, une telle ressource devrait permettre de mieux connaitre les phénomènes à l'œuvre dans le processus d'acquisition de l'écriture. D'un point de vue didactique, elle devrait appuyer la réflexion pédagogique des enseignants par des exemples réels de productions montrant ainsi les connaissances et les difficultés à chaque niveau. À terme, les développements TAL mis au point lors de la constitution du corpus devraient permettre, par exemple, le développement d'activités didactiques autour des difficultés repérées en corpus.

Dans le cadre global de ce projet, il nous semble nécessaire de maintenir des liens forts entre la recherche et le terrain. La constitution de ce corpus longitudinal sur cinq années nécessite un réseau d'enseignants impliqués et motivés. Dans cette optique, nous mettons à disposition de ce réseau, les corpus transcrits au fur et à mesure de leur conception sous forme de site web associé à certaines fonctions d'exploitation. Le site sur le corpus de CP sera ouvert dans les jours à venir et nous comptons sur les retours de ces utilisateurs pour affiner nos outils avant une mise à disposition globale auprès des enseignants et des chercheurs à la fin du projet.

Références

ANTONIADIS G., PONTON C., ZAMPA V. (2010). Exxelant et Mirto – Deux exemples d'environnement d'ALAO intégrant des outils TAL. *Multilinguisme et traitement des langues naturelles*. Montréal, Canada : PUQ.

AURIAC-SLUSARCZYK E., GUNNARSON-LARGY C. (2014). *Écriture et réécritures chez les élèves: un seul corpus, divers genres discursifs et méthodologies d'analyse*. Academia.

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

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

BLANCHARD A., KRAIF O., PONTON C. (2009). "Mastering Overdetection and Underdetection in Learner-Answer Processing: Simple Techniques for Analysis and Diagnosis". *Calico Journal*, *Vol. 26*(No. 3), 592-610.

DENIS P., SAGOT B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation* 4(46), 721-736.

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

FAIRON C., KLEIN J.R., PAUMIER S. (2006). *Le langage sms. Étude d'un corpus informatisé à partir de l'enquête « Faites don de vos sms à la science »*. Louvain-la-Neuve : Presses universitaires de Louvain.

GRANGER S., VANDEVENTER A., HAMEL M.-J. (2001). Analyse des corpus d'apprenants pour l'ELAO basée sur le TAL. *Traitement automatique des langues* 42(2), 609-621.

HABERT B., NAZARENKO A., SALEM A. (1997). *Les linguistiques de corpus*. Paris : Colin.

HEIFT T., SCHULZE M. (2007). *Errors and Intelligence in Computer-Assisted Language Learning: Parsers and Pedagogues*, New York and London : Routledge.

KENNEDY G. (1998). *An introducing to Corpus Linguistics*. London and New York : Longman.

KRAIF O., PONTON C. (2007). Du bruit, du silence et des ambiguïtés: que faire du TAL pour l'apprentissage des langues. Actes de *TALN* (Toulouse).

KUKICH K. (1992). *Techniques for Automatically Correcting Words in Text*. ACM Computing Surveys 24 (4), 377–439.

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

Wolfarth, C., Ponton, C., Totereau, C. (2016). Apports du TAL à la constitution et à l'exploitation d'un corpus scolaire. Corpus.