Aligner production et normalisation : une première approche pour l'étude d'écrits scolaires

Claire Wolfarth

Aligner production et normalisation : une première approche pour l'étude d'écrits scolaires

Claire Wolfarth[1]
(1) Univ. Grenoble Alpes, LIDILEM, F-38000 Grenoble, France
Claire.Wolfarth@univ-grenoble-alpes.fr

RESUME

L'émergence des corpus scolaires et la volonté d'outiller ces corpus spécifiques font apparaitre de nouvelles problématiques de recherche pour le traitement automatique des langues (TAL). Nous exposons ici une recherche qui vise le traitement de productions d'apprenants en début d'apprentissage de l'écriture, en vue d'une annotation et d'une exploitation ultérieure. À cette fin, nous proposons d'envisager cette étape comme une tâche d'alignement entre la production de l'apprenant et une normalisation produite manuellement. Ce procédé permet d'augmenter significativement les scores d'identification des formes et lemmes produits et améliore les perspectives d'annotation.

ABSTRACT

To align production and normalization : first approach to study school learner's writings
The emergence of school corpora and the will to provide tools for such specific corpora bring to light new research issues for natural language processing (NLP). We will expose a research which aims at correcting early learners' written production with the purpose of annotating and exploiting it at a later stage. We are putting forward to consider this stage as an alignment task between the learners' written and a manually produced 'normalized version' of it. This method increases significantly lemmas and forms identification task results and improves annotation possibilities.

MOTS-CLES : corpus scolaires, alignement, normalisation.
KEYWORDS: primary school corpus, alignment, normalization.

1 Production et normalisation : contexte de notre approche

Si l'étude de l'apprentissage de la lecture a permis l'outillage de son enseignement (Lector & Lectrix[1] ; Lectorino & Lectorinette[2] ; Bianco, 2017[3]), de tels outils n'existent pas encore pour l'apprentissage de l'écriture. À cette fin, plusieurs corpus scolaires de productions d'apprenants en langue française ont été constitués. Citons le corpus élaboré par M.-L. Elalouf (2005), rassemblant quelques 500 textes dans huit classes de CM2 et de 6e ; le corpus élaboré par H. Andersen, C.

[1] Cèbe, S., & Goigoux, R. (2009). Lector & [et] Lectrix: 20 posters pour l'étude des textes en collectif: CM1, CM2, 6e, Segpa. Retz.
[2] Goigoux, R., & Cèbe, S. (2013). Lectorino & Lectorinette CE1-CE2: apprendre à comprendre des textes narratifs. Retz.
[3] Bianco, M. (2017). Comment enseigner la compréhension ? *Enseigner à l'école primaire*. Hatier.

Leblay et E. Auriac-Slusarcyk (2010), contenant plusieurs centaines de récits et comptes-rendus scientifiques, recueillis en CE2, CM2, 6e et 4e ; le corpus recueilli par C. Garcia-Debanc et K. Bonnemaison (2014), de près de 400 textes autour d'une tâche de cohésion textuelle ; le corpus élaboré par J. David et C. Doquet (2016), rassemblant plus de 800 productions d'apprenants à l'école primaire et le corpus sur lequel nous travaillons qui contient près de 2 900 productions de textes, recueillis du CP au CE2 (Wolfarth *et al.*, 2017).

Au vu de la taille de certains de ces corpus, une exploitation manuelle est très coûteuse et difficilement envisageable. C'est pourquoi, nous cherchons à assister l'exploitation des corpus scolaires qui vise principalement à caractériser linguistiquement les écrits produits en cours d'apprentissage de l'écriture et à analyser les traces visibles de l'évolution de cet apprentissage. À ce titre, opérer un processus de reconnaissance des formes est nécessaire afin de permettre l'extraction du lexique et son analyse, l'analyse des erreurs et de phénomènes linguistiques spécifiques, etc.

Les spécificités de ces corpus et notamment leur grande distance à la norme (pour des exemples voir 2.1) nécessitent des outils spécifiques, inexistants à ce jour. Les outils tels que le *Trameur* (Fleury, 2007) permettent une exploitation textométrique des corpus (y compris scolaires) à condition que ceux-ci soient préalablement annotés. À l'heure actuelle, même si des travaux connexes étudiant l'apport du traitement automatique des langues (TAL) au domaine de l'apprentissage des langues existent (Granger *et al*., 2001 ; Antoniadis *et al*., 2010), peu se sont intéressés au traitement automatique d'écrits scolaires éloignés de la norme, et aucun outil ou méthode ne permet d'opérer les traitements nécessaires (reconnaissance des formes, extraction de lexique, annotation, etc.) de manière automatique ou semi-automatique.

Le travail présenté dans la suite de cet article, qui porte sur la normalisation et la reconnaissance des formes produites dans notre corpus, s'inscrit dans cet objectif et vise à outiller les corpus de textes scolaires. Mais il ne représente qu'une partie du processus que nous souhaitons mettre en œuvre et qui a pour but de proposer une aide à l'exploitation du corpus à des fins de description linguistiques. Il ne s'agit donc que d'une première étape, mais néanmoins nécessaire.

De plus, le travail que nous détaillons ici est un premier travail, réalisé uniquement avec des productions d'apprenants en fin de CP (première année d'apprentissage de l'écrit). Il conviendra donc, par la suite, d'étendre notre étude à l'ensemble du corpus. Cependant, nous pensons que les productions de niveau CP étant particulièrement complexes à traiter en raison du grand nombre d'erreurs présentes dans ces productions, nous pourrons nous inspirer de ces méthodes pour traiter les productions des années suivantes pour lesquelles on note notamment une amélioration de la compétence en segmentation en mots.

Dans la section suivante, nous présenterons des exemples de productions issus de notre corpus et discuterons de la notion de normalisation que nous utiliserons dans la suite de cet article. Puis, dans une troisième partie, nous reviendrons sur divers travaux réalisés autour du traitement de corpus peu normés. Enfin, nous présenterons dans la section 4, l'approche que nous mettons en œuvre pour la reconnaissance des formes dans les corpus d'écrits scolaires, avant de conclure sur les prolongements de notre travail dans la section 5.

2 Particularités des corpus scolaires

Dans cette partie, nous présenterons certaines caractéristiques des productions d'apprenants en fin de CP au travers d'exemples et différents choix que nous avons effectués pour notre corpus.

2.1 Exemples d'écrits produits en fin de CP

Pour chaque exemple présenté dans cette partie, nous préciserons ses particularités et les types d'écarts à la norme qu'ils contiennent et qui nous intéressent plus particulièrement.

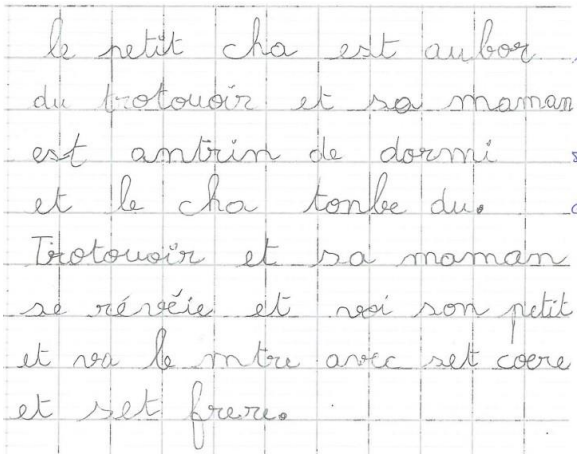


FIGURE 1 : Production de l'élève 96 en fin de CP

Version orthographiée selon la norme[4] : « Le petit chat est au bord du trottoir et sa maman est en train de dormir et le chat tombe du trottoir et sa maman se réveille et voit son petit et va le mettre avec ses sœurs et ses frères. »

Cet exemple illustre une partie des écarts à la norme que produisent les apprenants en fin de CP. On relève en effet la présence de variantes orthographiques, c'est-à-dire de formes graphiques non normées d'un mot, comme les formes *trottoir* et *sœurs*, respectivement orthographiées « trotouoir » et « coere ». On note également des écarts de segmentation en mots, et notamment un cas de sous-ponctuation : « antrin » (*en train*) et des écarts de segmentation en phrase : « tonbe du. Trotouoir » (sur-segmentation). Sur cet exemple, produire une version normée d'un point de vue orthographique est relativement aisé, mais ce n'est pas le cas de toutes les productions : observons l'exemple suivant.

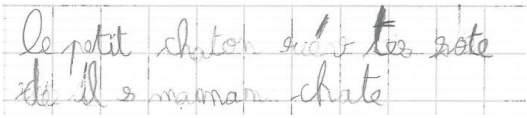


FIGURE 2 : Production de l'élève 2453 en fin de CP

[4] Nous empruntons ici l'appellation « Version orthographiée selon la norme » à C. Boré et M.-L. Elalouf (2017).

Pour cet exemple, la tâche de réécrire la production en normant l'orthographe est davantage une tache d'interprétation que de réécriture. La tâche qui consiste à produire une version orthographiée selon la norme produit donc une version incomplète « le petit chaton se réveille et saute de […] maman chatte ».

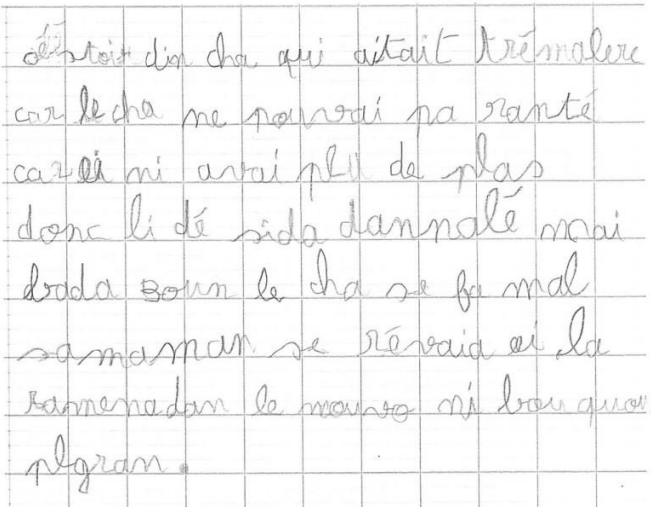


FIGURE 3 : Production de l'élève 2972 en fin de CP

Version orthographiée selon la norme : « C'est l'histoire d'un chat qui était très malheureux car le chat ne pouvait pas rentrer car il n'y avait plus de place donc il décida de s'en aller mais badaboum le chat se fait mal <segmentation/> sa maman se réveilla et le ramena dans le nouveau nid beaucoup plus grand. »

Cet exemple présente également divers types d'écart à la norme. On constate ainsi de nombreuses variantes orthographiques comme « révaia » (*réveilla*), « plas » (*place*) ou encore « malere » (*malheureux*). Cette production présente également des écarts de segmentation en mots, comme dans la séquence *beaucoup plus grand*, produite « bou quou plgran », que ce soit des cas de sous-segmentation : « trémalere » (*très malheureux*), « plgran » (*plus grand*), etc. ou de sur-segmentation : « dé sida » (*décida*). On note également un cas d'absence de marqueur de segmentation en phrases (que ce soit un signe de ponctuation ou un connecteur comme *et* ou *mais*), marqué par la balise <segmentation/> dans la proposition de version orthographiée selon la norme.

Les écarts relevés dans ces productions ne sont que des exemples d'écarts à la norme attestés dans les productions de CP, d'autres auraient pu être mentionnés comme les problèmes d'accord, le non-respect des concordances de temps ou encore des tournures syntaxiques spécifiques mais dans cette première approche des corpus scolaires, nous nous sommes principalement focalisés sur la segmentation en mots et sur les aspects orthographiques.

2.2 Intérêt et construction de la normalisation

Nous proposons d'entamer une réflexion en vue d'une aide à l'annotation et à l'exploitation des corpus scolaires, en adoptant une démarche analogue à celle d'O. Kraif et de C. Ponton (2007). Afin

d'envisager l'ensemble des phénomènes décrits dans le paragraphe précédent dans le traitement et l'exploitation des productions d'apprenants, nous proposons une approche s'appuyant à la fois sur une transcription des productions et sur une normalisation de celles-ci. Nous faisons l'hypothèse qu'un alignement forme à forme entre transcriptions et normalisations permettra l'identification des formes produites dans les productions et facilitera donc l'exploitation ultérieure du corpus.

Dans notre corpus, nous disposons donc de trois versions de chaque production :

- Un *scan* de la production manuscrite de l'élève ;
- Une *transcription* tapuscrite réalisée manuellement qui reproduit le contenu linguistique de la production[5]. Précisons que cette étape est réalisée manuellement, les systèmes d'OCR actuels n'étant pas adaptés au type d'écrit de notre corpus ;
- Une *normalisation*.

Si la normalisation utilisée par notre approche n'est pas à construire pour les productions issues d'épreuves de dictées, elle reste à définir pour les productions de textes. Nous pensons que cette approche est généralisable à d'autres formes de productions comme celles issues d'exercices de reformulation ou de copie. Un alignement entre la transcription des productions d'apprenants et leur normalisation devrait alors permettre des allers/retours entre ces deux versions afin de faciliter des interrogations fines de ces corpus, portant par exemple sur le temps des verbes produits par les apprenants, les constructions syntaxiques utilisées. Ce processus permettra également de mesurer les écarts entre formes produites et formes normées afin d'identifier les erreurs les plus fréquentes, ou encore l'évolution de ces erreurs.

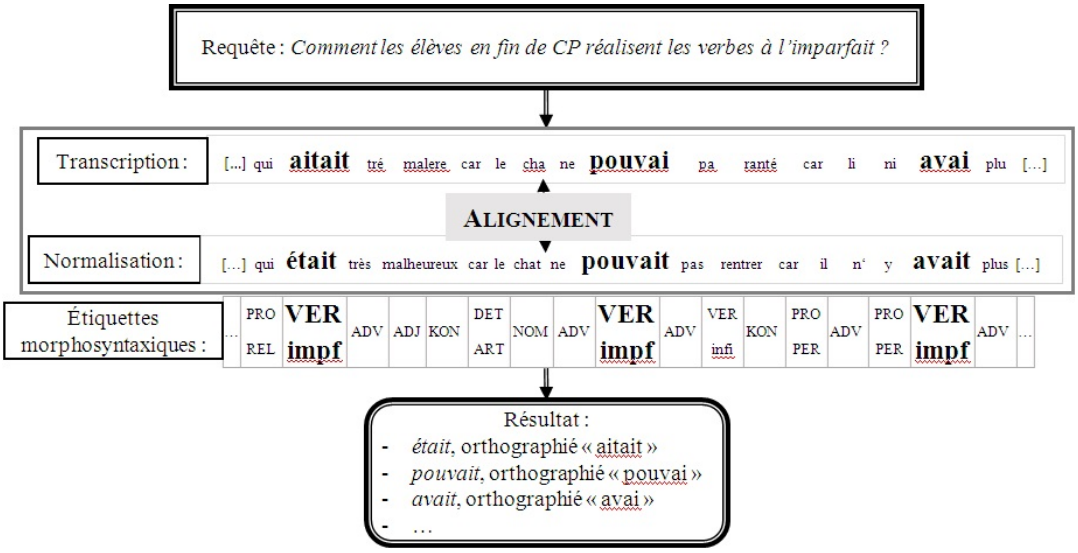


FIGURE 4 : Exemple de fonctionnement de l'outil et de l'usage de l'algorithme d'alignement à l'aide d'une requête et de la production 2972

La figure 4 permet d'illustrer un exemple de requête possible grâce à une approche par alignement entre transcriptions et normalisations. Nous donnons ici l'exemple d'une recherche de verbes à l'imparfait. Un étiquetage morphosyntaxique de la normalisation grâce à des outils de TAL déjà existants (TreeTagger dans cet exemple) permettra de retrouver les verbes à l'imparfait sous leur

[5] Pour plus de détails sur cette étape de transcription, voir Wolfarth et al., 2017.

forme normée, tandis que l'algorithme d'alignement permettra de retrouver la forme graphique attestée en corpus de ces verbes. Le résultat d'une telle requête peut être étudié sous différents aspects : usage de l'imparfait dans les productions d'apprenants, réalisation de la finale des verbes en *–ait*, respect ou non de la phonologie dans le cas des erreurs de morphologie verbale, etc.

Afin de réaliser l'alignement, cette transcription est alors comparée à une version normée de la production. Celle-ci doit permettre l'identification des phénomènes que l'on souhaite analyser ; les choix que l'on va effectuer lors de cette étape sont donc essentiels. En effet, tout phénomène non normé ne pourra être retrouvé à l'aide des outils de traitement automatique. Par exemple, si l'on souhaite analyser toutes les formes produites du mot *chat*, il va être nécessaire de normer son orthographe. Si ce cas semble relativement évident, nous avons dû faire des choix bien moins aisés. Considérons l'exemple de la production 1280 (FIGURE 5) : dans la séquence « sa maman réquonsili le bébé chat », il semble que l'enfant a confondu le verbe *réconcilier* et le verbe *réconforter*. Il faudrait normaliser à l'aide du verbe *réconforter*. Mais en procédant à cette substitution, il parait difficile de retrouver le verbe initial dans la transcription puisqu'il n'est pas correctement orthographié et donc non identifiable automatiquement. Dans ce cas, nous avons donc choisi de garder le verbe *réconcilier* et de ne pas le remplacer par le verbe attendu *réconforter*. L'étape de normalisation a fait émerger de nombreuses réflexions similaires.

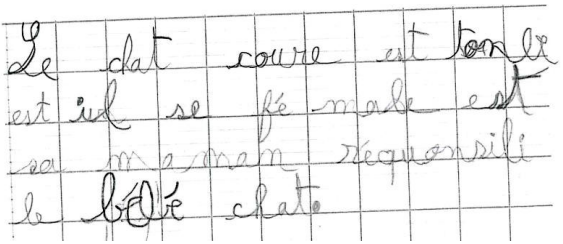


FIGURE 5 : Production de l'élève 2972 en fin de CP

Version orthographiée selon la norme : « Le chat court et tombe et il se fait mal et sa maman réconcilie le bébé chat. »

Dans la plupart des cas, le choix réalisé a été de rester au plus proche de la production de l'apprenant, à l'exception de l'orthographe, de la segmentation en mots et des accords, qui ont été normés. Nous avons également fait le choix de marquer, à l'aide d'une balise, l'absence de segmentation en phrase. La normalisation présente donc une structure syntaxique et lexicale proche de la production initiale.

Les choix que nous présentons ici ont complexes et sont ceux qui ont été adoptés pour calculer les résultats présentés dans la suite de cet article. Il est à noter que ces choix sont amenés à évoluer.

3 Travaux connexes

Les travaux de normalisation (selon le sens en vigueur dans le domaine du traitement automatique des langues) et de correction ne sont pas nouveaux en TAL et ont été envisagés dans de nombreux domaines. On peut citer notamment les travaux qui ont cours pour les corpus de SMS, de tweets ou d'écrits issus des réseaux sociaux et des forums en lignes (Kobus *et al*., 2008 ; Beaufort *et al*., 2010 ; Kogkitsidou, 2016). De nombreux travaux ont également vu le jour dans le domaine de l'Apprentissage des Langues Assisté par Ordinateur (ALAO) et de l'apprentissage des langues

secondes (Granger *et al.*, 1998 ; Antoniadis *et al.*, 2005 ; Antonsen, 2012). Enfin, on mentionnera également de nombreux travaux dans le domaine de la correction, orthographique le plus souvent, pour scripteurs experts. Mais, à ce jour, peu de travaux concernent l'apprentissage de l'écrit et la normalisation ou la correction de productions d'apprenants de l'écriture.

Dans le cadre de ces différents travaux, diverses approches ont été avancées. Nous présenterons dans cette partie quelques-unes de ces approches. Citons, en premier lieu, les approches par lexiques. Dans ces approches, les formes produites jugées incorrectes ou non normées sont comparées à des formes issues de lexiques selon des indices graphiques, que ce soit à partir des distances d'édition proposées par V. Levenshtein (Kernighan *et al.,* 1990) ou à partir de clés de similarité graphiques (Ndiaye, Vandeventer Faltim, 2004), ou des indices phonologiques (Belrhali, 1995).

Cette approche a été testée sur notre corpus (Wolfarth, 2017) au moyen de comparaisons avec la base lexicale MANULEX (Ortéga, Lété, 2010) basées sur des indices phonologiques, mais la trop grande ambiguïté des résultats de cette approche et la difficulté de la désambiguïsation en raison de la présence de structures syntaxique non encore maitrisées nous poussent à expérimenter une autre méthode.

Certains travaux font également état d'approches par génération d'erreurs (Cohard, 1988 ; Antonsen, 2012). Dans ces approches, de nombreuses formes erronées sont générées à l'aide de règles linguistiques ou d'automates à états finis puis comparées aux formes du corpus en présence. Ces approches sont particulièrement adaptées au traitement des erreurs de morphologie ou aux corpus présentant des règles récurrentes comme les corpus de L2 produits par des locuteurs issus des mêmes L1, mais ne conviennent pas aux productions de CP, comportant un trop grand nombre d'erreurs non régulières.

Enfin, et c'est cette approche qui va nous intéresser dans la suite de cet article, certains travaux envisagent les tâches de correction ou de normalisation comme une tâche d'alignement entre la production attestée en corpus et la production attendue. Cette approche a notamment été abordée dans le cadre de la correction de dictées.

C. Santiago-Oriola (1998) a ainsi conduit une recherche dans une perspective d'aide à la correction de dictées. Dans ce travail, la détection des erreurs est réalisée non pas à l'aide d'un correcteur orthographique mais à l'aide d'un alignement de la dictée et de la production de l'apprenant s'appuyant sur des indices phonologiques et des règles de transformation phonologique.

Plus récemment, R. Beaufort et S. Roekhaut (2011) ce sont appuyés sur un algorithme d'alignement dans la perspective d'automatiser la correction d'exercices de dictées. Il ne s'agissait plus alors de s'appuyer sur des indices phonologiques mais sur une analyse morphosyntaxique automatique de l'original de la dictée. L'alignement est alors opéré grâce à des indices graphiques, issus des opérations d'éditions classiques et de substitutions de séquences graphiques fréquentes au cours de l'apprentissage du français.

Une méthode similaire a été proposée pour un exercice de traduction (Desmet, Hérogue, 2005). Dans ce travail, plusieurs traductions possibles sont générées en guise de version normée. La traduction retenue pour l'alignement est alors la traduction ayant le plus de similitudes avec la réponse de l'apprenant.

Nous proposons ici de tester une approche similaire pour le traitement d'écrits scolaires.

4 Alignement par indices phonologiques

Étant donné le choix effectué lors de la normalisation des productions d'apprenants de modifier le moins possible ces productions et le peu de modifications syntaxiques apportées, nous pouvons nous appuyer sur un alignement linéaire des deux versions d'une production. De plus, l'objectif de notre démarche d'alignement s'inscrivant dans un objectif ultérieur d'annotation des formes, l'alignement développé est un alignement de formes, c'est-à-dire de tokens. Toutefois, lorsque cela sera nécessaire à l'annotation, c'est-à-dire lorsque la forme produite et la forme normée ne seront pas identiques, un alignement ultérieur, partiel ou total, de graphèmes, proche de ce que propose C. Santiago-Oriola, pourra être envisagé lors de l'étape d'annotation mais ne concerne pas les propos que nous développons ici.

L'algorithme d'alignement présenté ici a été développé à partir d'un corpus de référence contenant 50 productions de fin de CP, choisies aléatoirement parmi les textes des élèves pour lesquels nous disposions d'un suivi longitudinal et alignées manuellement. À terme, l'aligneur développé devra être adapté à l'ensemble des niveaux de primaires (CP à CM2), il conviendra donc de développer un véritable corpus aligné de référence.

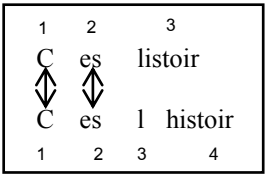
L'aligneur que nous avons développé et que nous présentons par la suite n'est qu'une première version dont l'objectif est double. D'une part, il s'agit de montrer la faisabilité de l'approche et, d'autre part, de disposer d'un premier outil nous permettant d'avancer sur l'exploitation du corpus qui représente le cœur de notre travail. Il est particulièrement conçu pour prendre en considération deux caractéristiques principales des corpus d'apprenants en début d'apprentissage de l'écriture. En effet, il prend en compte les erreurs de segmentations en mots. Cet aspect sera développé dans le paragraphe 4.1. Mais il tient également compte de la dimension phonologique des écrits d'apprenants et s'appuie sur celle-ci, que nous le verrons dans le paragraphe 4.2.

4.1 Principe de l'alignement

L'algorithme développé est un algorithme itératif qui considère chaque forme les unes après les autres et ne s'interrompt qu'une fois l'ensemble des formes de chacune des versions examinées. Le principe de l'algorithme est détaillé ci-dessous.

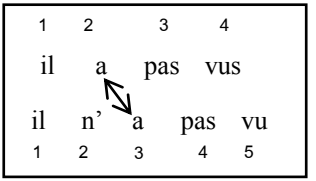
1. Dans le cas le plus courant, l'algorithme procède à une comparaison des formes de rang *n*, selon les mesures exposées au paragraphe suivant (section 4.2). Si les formes sont jugées équivalentes, elles sont alignées, ce sera également le cas pour les étapes suivantes.

```
POUR (n de 0 à longueur(Normalisation)) {
  SI (COMPARER(Transcription[n],Normalisation[n]) == TRUE) {
    ENREGISTRER(Transcription[n],Normalisation[n]) ;
  }
}
```



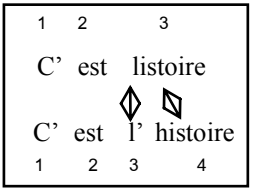
2. Bien que peu de modifications syntaxiques aient été apportées à la normalisation, il se peut que celle-ci comporte tout de même quelques suppressions ou ajouts. L'algorithme permet donc des comparaisons entre la forme produite de rang *n* et la forme normée de rang *n+1*, *n+2* ou *n+3* (cas d'ajout) et inversement (cas de suppression).

```
POUR (n de 0 à longueur(Normalisation) {
  POUR (i de 0 à 3) {
    SI (COMPARER(Transcription[n],Normalisation[n+i]) == TRUE) {
      ENREGISTRER(Transcription[n],Normalisation[n+i]) ;
    }
    SI (COMPARER(Transcription[n+i],Normalisation[n]) == TRUE) {
      ENREGISTRER(Transcription[n+i],Normalisation[n]) ;
    }
  }
}
```



3. Afin de gérer les erreurs de sous-segmentation, comme « sereveille » (*se réveille*), et de sur-segmentation, comme « dé sida » (*décida*), nous avons également dû implémenter ce que M. Jansche (2001) appelle une *fenêtre glissante* et qui envisage la comparaison de plusieurs formes produites à une forme normée (cas de sur-segmentation) et inversement (cas de sous-segmentation). L'algorithme ainsi développé autorise une fenêtre de longueur 3, considérant qu'une forme produite donnée peut correspondre à trois formes normées et inversement. Cette longueur a été choisie à partir des observations manuelles réalisées au préalable, mais des tests à l'aide de longueurs différentes sont à mener pour déterminer la longueur de fenêtre la plus efficiente.

```
POUR (n de 0 à longueur(Normalisation) {
  SI (COMPARER(Transcription[n], Normalisation[n].Normalisation[n+1]) == TRUE) {
    ENREGISTRER(Transcription[n], Normalisation[n].Normalisation[n+1]) ;
  }
  SI (COMPARER(Transcription[n], Normalisation[n].Normalisation[n+1].Normalisation[n+2]) ==
    TRUE) {
    ENREGISTRER(Transcription[n], Normalisation[n].Normalisation[n+1] .Normalisation[n+2]) ;
  }
  SI (COMPARER(Transcription[n].Transcription[n+1],Normalisation[n]) == TRUE) {
    ENREGISTRER(Transcription[n].Transcription[n+1],Normalisation[n]) ;
  }
  SI (COMPARER(Transcription[n].Transcription[n+1].Transcription[n+2],Normalisation[n]) ==
      TRUE) {
    ENREGISTRER(Transcription[n].Transcription[n+1].Transcription[n+2],Normalisation[n]) ;
  }
}
```



4. Enfin, l'algorithme se déployant de façon linéaire, il permet d'aligner les segments restants, qui ne peuvent être alignés par les mesures explicitées ici.

Si ces étapes permettent la comparaison de plusieurs formes entre elles, chaque comparaison est effectuée selon différentes mesures afin d'améliorer les possibilités d'identifier les équivalences.

4.2 Mesures de comparaison

Les travaux de C. Santiago-Oriola nous intéressent particulièrement puisque, partant du constat établi précédemment (Wolfarth, 2017) que seules 25 % des erreurs commises par les apprenants en fin de CP entrainent des modifications d'ordre phonologique, nous choisissons de nous baser sur des

indices de type phonologique pour construire un algorithme d'alignement. Comme nous l'avons mentionné, nous utilisons un algorithme d'alignement linéaire. Celui-ci considère deux formes de rang équivalent, une forme issue de la production de l'apprenant, l'autre issue de la production normée, et les compare selon les modalités suivantes :

1. Comparaison graphique stricte : cette première comparaison permet de repérer les formes correctement orthographiées par les apprenants. Elles sont considérées ainsi lorsqu'il y a correspondance exacte entre forme normée et forme produite;

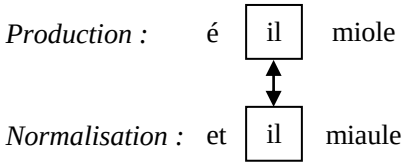


FIGURE 6 : Résultat de la comparaison graphique stricte

2. Comparaison phonologique : il s'agit ici de comparer les représentations phonologiques des formes. Celles-ci sont obtenues à l'aide du module LIA_PHON (Béchet, 2001). Ce module permet de convertir n'importe quelle forme graphique en suite de phonèmes, que cette forme soit ou non correctement orthographiée. La représentation phonologique est produite en suivant les règles de prononciation du français ;

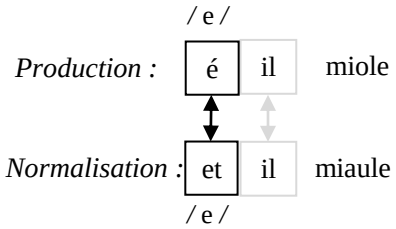


FIGURE 7 : Résultat de la comparaison phonologique stricte

3. Comparaison phonologique avec relâchement de contraintes (ou comparaison archi-phonologique) : nous adoptons ici la notion d'archiphonème, développée par N. Catach (1995). Un archiphonème est un « représentant de l'ensemble des traits phoniques pertinents communs à deux ou plusieurs phonèmes, qui sont par rapport aux autres dans un rapport exclusif » (p.16). Nous regroupons par exemple les phonèmes /ø/, /œ/ et /ə/ sous l'archiphonème /Œ/. Ce choix fait suite au constat que les apprenants en fin de CP ne maitrisent pas toujours la distinction entre les phonèmes /ø/, /œ/ et /ə/ ; /e/ et /ɛ/ ; /o/ et /ɔ/ ; /ɛ̃/ et /œ̃/, d'autant plus que les frontières entre ces phonèmes peuvent être différentes d'une région à une autre et qu'ils peuvent être transcrits par un même phonème. Citons pour exemple, les formes *comme* et *comment* où le graphème *o* transcrit à la fois les phonèmes /o/ et /ɔ/. Il s'agit donc ici de comparer les représentations phonologiques des formes produites par LIA_PHON et dont les phonèmes ont été convertis en archiphonèmes

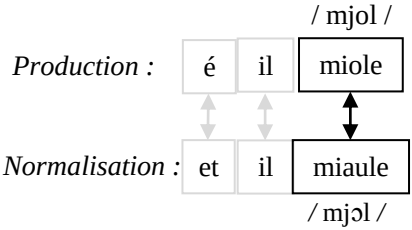


FIGURE 8 : Résultat de la comparaison phonologique avec relâchement de contraintes

Une quatrième voie de comparaison devrait bientôt voir le jour afin de permettre un relâchement des contraintes encore plus étendu grâce à la prise en compte des différentes voies de phonétisation des graphèmes selon leur contexte. Prenons un exemple : la suite de trois lettres *e-s-t* peut être lue /ɛ/ lorsqu'il s'agit du verbe *être* ou /ɛst/ lorsqu'elle est présente dans l'unité lexicale *estimer* ou le point cardinal *est*. C'est cette multiplicité des voies de phonétisation qui amène un enfant de CP qui veut écrire la forme *énorme* à l'orthographier « estnorme ». Dans ce cas, le module LIA_PHON qui suit les règles de lecture du français identifiera la suite phonémique/ɛstnɔrmø/, une prise en compte des différentes possibilités de prononciation d'un graphème ou d'une suite de graphèmes va donc être nécessaire pour retrouver ici la forme normée.

5 Résultats et discussion

Cet algorithme d'alignement ayant été élaboré, dans un premier temps, à partir de productions de CP, il a été testé sur un échantillon de 50 productions de fin de CP, soit 1 440 formes. Pour en mesurer l'efficience, nous employons les scores suivants : la précision, le rappel et la F-mesure. Le corpus de référence à partir duquel ces scores sont évalués a été aligné à la main par un seul annotateur.

Ces scores ont été calculés pour chacune des mesures de comparaison présentées au paragraphe 4.2 et sont résumés dans le tableau suivant.

Mesure	Alignement par comparaison graphique stricte	Alignement par comparaison phonologique	Alignement par comparaison phonologique avec relâchement de contraintes
Précision	83,3 %	86,3 %	86,7 %
Rappel	59,9 %	72,6 %	74 %
F-mesure	69,7 %	78,9 %	80 %

TABLEAU 1 : Scores de l'algorithme selon les indices de comparaison utilisés

Les scores de l'algorithme incluant la prise en compte de toutes les règles de prononciation du français n'ont pas encore pu être calculés mais intégrer cette dernière mesure de comparaison devrait permettre d'améliorer quelque peu le score de l'algorithme. Au vu des résultats, il apparait qu'intégrer des indices phonologiques à notre algorithme permet d'améliorer le score de F-mesure de plus de 10 %.

Mentionnons cependant qu'une partie des erreurs provient de la présence de productions telles que les productions 2453 (FIGURE 2) et 208 (FIGURE 10), particulièrement éloignée de la norme, pour lesquelles une interprétation humaine et fort peu aisée. Un traitement automatique de ces productions semble donc proscrit.

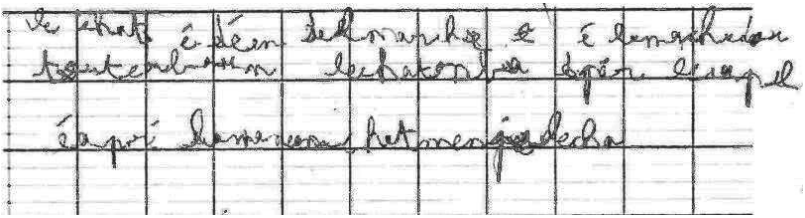


FIGURE 9 : Production de l'élève 208 en fin de CP

Version orthographiée selon la norme : « Le chat est descendu de [la|il][6] marche et la marche […]. Tout à coup boum <segmentation/> le chat tomba et pleure <segmentation/> [le chat|il] appelle et après la mère ramène <segmentation/> le chat mange le chat. »

Cette méthode semble apporter des résultats très intéressants pour notre corpus. Cependant, l'algorithme utilisé dans cet article est un algorithme original, conçu spécifiquement pour le traitement des productions d'apprenant. Il conviendrait donc d'évaluer les approches classiques d'alignement pour les comparer aux résultats produits par cet algorithme.

Évidemment, l'approche par alignement est coûteux puisqu'elle nécessite de produire une normalisation manuellement. La taille des corpus scolaires ne permet cependant pas d'envisager l'emploi d'un apprentissage automatique. Néanmoins, ce premier travail permettra peut-être de produire suffisamment de données d'alignement pour permettre une méthode par apprentissage dans le cas d'un nouveau corpus produit dans un contexte similaire.

Ce travail a également montré la pertinence d'utiliser des indices phonologiques dans le cas de l'apprentissage de l'écrit par des locuteurs natifs. Il pourrait donc être intéressant de mesurer la performance d'un outil spécifique de normalisation inspiré des méthodes de reconnaissance vocale, comme cela a été fait dans certains travaux s'intéressant au traitement des SMS (Kobus *et al*., 2008). Il conviendrait également d'utiliser d'autres outils de phonétisation afin de comparer leur efficience.

Enfin, cette étude a été réalisée à partir de productions de CP, et nécessitera vraisemblablement des adaptations si l'on s'intéresse à des productions de CE1 et au-delà, au vu des grandes évolutions qui apparaissent dans les premières années d'apprentissage de l'écriture.

Remerciements

Un grand merci à Claude Ponton et Catherine Brissaud, qui encadrent cette thèse, pour leurs relectures attentives.

[6] Les symboles [x_1|x_2] marquent le doute entre x_1 et x_2.

Références

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.

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.

ANTONSEN L. (2012). Improving feedback on L2 misspellings-an FST approach. Actes de *Proceedings of the SLTC 2012 workshop on NLP for CALL*; Lund; 25th October; 2012 (No. 080, pp. 1-10). Linköping University Electronic Press.

BEAUFORT R., ROEKHAUT S., COUGNON L. A., & FAIRON C. (2010). Une approche hybride traduction/correction pour la normalisation des SMS. Actes de *Actes de la 17e conférence sur le traitement automatique des langues naturelles (TALN'10)* (pp. 19-23).

BEAUFORT R., ROEKHAUT S. (2011). Automation of dictation exercises. *Computational Linguistics in the Netherlands Journal*, (1), 1-20.

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

BELRHALI, R. (1995) : Phonétisation automatique d'un lexique général du français : systématique et émergence linguistique. Thèse. Université Stendhal, Grenoble.

BORE C., ELALOUF M.-L. (2017). Deux étapes dans la construction de corpus scolaires : problèmes récurrents et perspectives nouvelles. In Doquet C., David J., Fleury S., *Spécificités et contraintes des grands corpus de textes scolaires : problèmes de transcription, d'annotation et de traitement*, Corpus, 16 | déc. 2016, 185-214.

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 Gruaz C. et Duprez D.). Paris : Nathan.

COHARD B. (1988). Logiciel de détection et de correction des erreurs lexicales. Mémoire de master. CNAM.

DAVID J., DOQUET C. (2016). Les écrits d'élèves: un corpus de référence pour le français contemporain. Actes de *SHS Web of Conferences* (Vol. 27, p. 11001). EDP Sciences.

DESMET P., HEROGUEL A. (2005). Les enjeux de la création d'un environnement d'apprentissage électronique axé sur la compréhension orale à l'aide du système auteur IDIOMA-TIC. Actes d *Alsic. Apprentissage des Langues et Systèmes d'Information et de Communication*, 8(1), 281–303.

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

FLEURY S. (2007). Le Trameur, Manuel d'utilisation. - http://tal.univ-paris3.fr/trameur/leMetierLexicometrique.pdf. 2007

GARCIA-DEBANC C., BONNEMAISON K. (2014). La gestion de la cohésion textuelle par des élèves de 11-12 ans: réussites et difficultés. Actes de *SHS Web of Conferences* (Vol. 8, pp. 961-976). EDP Sciences.

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.

KERNIGHAN M. D., CHURCH K. W., GALE W. A. (1990). A spelling correction program based on a noisy channel model. Actes de *Proceedings of the 13th conference on Computational linguistics*-Volume 2 (pp. 205-210). Association for Computational Linguistics.

KOBUS C., YVON F., DAMNATI G. (2008). Normalizing SMS: are two metaphors better than one? Actes de *Proceedings of the 22nd International Conference on Computational Linguistics*(1), Manchester, UK, 441–448.

KOGKITSIDOU E., ANTONIADIS G.. (2016) L'architecture d'un modèle hybride pour la normalisation de SMS. Actes de *la 23e conférence sur le traitement automatique des langues naturelles (TALN'16)* (pp.355-363).

KRAIF O., PONTON C. (2007). Du bruit, du silence et des ambiguïtés: que faire du TAL pour l'apprentissage des langues. Actes de *la 14e conférence sur le traitement automatique des langues naturelles (TALN'07).*

NDIAYE M. ET VANDEVENTER FALTIN A. (2004). Correcteur orthographique adapté à l'apprentissage du français. BULAG, 29:117–134.

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

SANTIAGO ORIOLA C. (1998). Système vocal interactif pour l'apprentissage des langues. La synthèse de la parole au service de la dictée. Thèse. Toulouse III.

WOLFARTH C., PONTON C., BRISSAUD C. (2016) Du TAL dans les écrits scolaires: premières approches. Actes de *RECITAL'16 - rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues.*

WOLFARTH C., PONTON C., TOTEREAU C. (2017). 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 16*, janv. 2017, 185-214.