# Les écrits d'élèves : un corpus de référence pour le français contemporain

Jacques David, Claire Doquet

# Les écrits d'élèves : un corpus de référence pour le français contemporain

David, Jacques

Université de Cergy-Pontoise
Laboratoires AGORA – EA 7392
& CLESTHIA – EA 7345

Doquet, Claire

Université La Sorbonne nouvelle – Paris 3
Laboratoire CLESTHIA – EA 7345

**Résumé :** La situation d'apprentissage de l'écriture et les difficultés qu'elle révèle mettent au jour des zones de la langue qui apparaissent caractéristiques et délicates à maîtriser chez les scripteurs débutants. Ce constat nous a conduits à constituer une base de données d'écrits d'élèves en vue d'explorer systématiquement les différentes composantes du français qui résistent ou qui évoluent dans le processus d'acquisition de l'écriture. Notre communication retrace, dans un premier temps, les problèmes spécifiques posés par ce type de corpus, en particulier pour ce qui concerne le protocole de transcription et d'annotation. Les écarts à la norme des apprentis scripteurs doivent en effet faire l'objet d'annotations spécifiques pour que les éléments verbaux soient lemmatisés correctement. Dans une deuxième partie, nous exposons deux types d'investigations menées dans ce corpus à propos de la ponctuation blanche : i) les espaces et blancs de textes ; ii ) la segmentation de mots. Il s'agit de montrer comment le corpus que nous constituons, à travers les conventions d'annotation qui ont été construites, permet de repérer des éléments caractéristiques de la manière dont les élèves délimitent et ponctuent leurs écrits, mais aussi d'observer comment cette ponctuation blanche évolue pour en extraire des indications propres à accompagner ou renouveler les apprentissages induits.

## 1. Introduction

Nous exposons dans cet article un travail en cours, entamé en 2013, de constitution d'une base de données d'écrits d'élèves et d'un outillage technologique et linguistique d'exploration. Cette base de données, longitudinale (les écrits recueillis, au nombre actuel de 1250, couvrent la scolarité primaire et secondaire), comportera pour chaque texte son avant-texte (plan, notes, brouillon, etc.) ainsi que des métadonnées de types sociologique et didactique qui permettront d'interroger les données de manière sélective ; le logiciel d'analyse textuelle associé, Le Trameur1, sera augmenté de modules spécifiques pour l'analyse de ces écrits particuliers que constituent les productions écrites à différents stades de l'apprentissage ; le corpus sera mis à disposition sur Ortolang et restera ouvert pour recevoir par la suite de nouveaux écrits.

La plupart des éléments présentés ici sont le fruit du travail de notre groupe de recherche. Concernant les éléments de transcription et d'annotation, que nous évoquons en première partie, ils proviennent d'un travail collectif entre plusieurs laboratoires : EA 7345 Clesthia, UMR 5263 CLEE ERSS, EA 609 Lidilem, EA 4507 EMA, EA 4671 ADEF.

Le choix de travailler sur les écrits des élèves est lié, du point de vue de la linguistique, à la possibilité d'un retour des observations vers la description du système de la langue. La situation d'apprentissage de l'écriture et les difficultés qu'elle révèle permettent de mettre au jour les composantes les plus résistantes

[1] Le Trameur est un logiciel d'analyse textuelle conçu par Serge Fleury dans le cadre du laboratoire Clesthia (EA 7345). http://www.tal.univ-paris3.fr/trameur/

de la langue qui apparaissent aussi, mais sous forme atténuée, chez les scripteurs disposant d'un haut degré de maîtrise de l'écrit. Explorer ces écrits en grands corpus permet de repérer des régularités et, à partir de l'observation statistique, de déterminer quels sont les aspects du système linguistique dont l'acquisition est la plus complexe, puis de mieux comprendre la manière dont les scripteurs investissent le système pour s'en approprier les règles. Les spécificités des écrits des élèves interrogent également le travail technique d'aménagement des logiciels. Il est ainsi envisager de pouvoir repérer automatiquement les erreurs et les dysfonctionnements afin, d'une part, de faciliter le travail d'annotation et, d'autre part, d'améliorer l'analyse textométrique à partir d'indicateurs linguistiques fiables.

## 2. Les spécificités des corpus d'écrits d'élèves et leurs conséquences en termes de transcription / annotation

Pour rendre accessibles les écrits des élèves que nous avons rassemblés, une première opération de transcription est nécessaire : elle permet de faciliter la lecture et de désambigüiser certaines graphies. Cette transcription, destinée à être lue par des utilisateurs de notre base de données, est suivie de l'annotation manuelle de chaque texte, base de production du fichier qui sera exploité par le logiciel d'analyse textuelle. Ces deux opérations, que nous traitons successivement ici, sont fortement impactées par deux caractéristiques des textes : l'existence pour un même texte de plusieurs versions et souvent l'intervention d'au moins deux scripteurs d'une part, et d'autre part les écarts à la norme langagière que présentent les textes des élèves.

### 1.1 Les différents états des textes et les commentaires des enseignants

Une des injonctions récurrentes des textes officiels sur l'enseignement-apprentissage de l'écriture est aujourd'hui la nécessité d'accompagner la réécriture, tâche courante chez les écrivains qui apparaît comme devant être enseignée aux élèves. Les productions d'écrits font très souvent l'objet d'au moins une réécriture (Doquet, 2011), qui fournit deux états de chaque texte : la première et la deuxième version, que nous nommerons par commodité le *brouillon* et le *texte final*. Entre ces deux versions d'un même texte, on observe souvent des interventions de l'enseignant qui lit et commente le brouillon avant que ce dernier ne soit repris par l'élève et réécrit. Dans le souci de mieux comprendre le processus de l'écriture, il est bien entendu indispensable de tenir compte des deux versions du texte mais aussi des commentaires de l'enseignant, qui induisent souvent fortement la réécriture[2].

Le choix a été fait, pour la transcription, de se conformer aux principes de la transcription diplomatique, que F. Masai (1950) qualifiait de « relevé archéologique des textes, tels qu'ils sont transmis par les manuscrits existants [… qui] a pour mission de reproduire fidèlement le document tel qu'il est sorti de l'officine productrice » (pp.185 et 187). De la transcription diplomatique philologique telle que défendue par F. Masai, à celle que pratique la génétique textuelle (Grésillon, 1994), si les objets de recherche changent, la technique demeure, même si les travaux contemporains relativisent la « fidélité » de la reproduction des manuscrits : d'après A. Crasson et J.D. Fedeke (2007), « la transcription diplomatique *photographie* le document en rapportant, avec les outils qui le permettent, malgré leurs limites, tous les événements du manuscrit ». Par *événement du manuscrit*, il faut entendre l'ensemble des traces que laisse l'activité d'écriture, y compris par exemple des dessins ou la couleur de l'encre, mais aussi, bien entendu, les ratures : biffures, segments textuels hors ligne ou en marge, qui permettent au lecteur de reconstituer des opérations scripturales : l'ajout, la suppression et leurs composés, le remplacement et le déplacement (Grésillon, 1994).

Les conventions de transcription de ces opérations majoritairement[3] suivies à l'Institut des Textes et

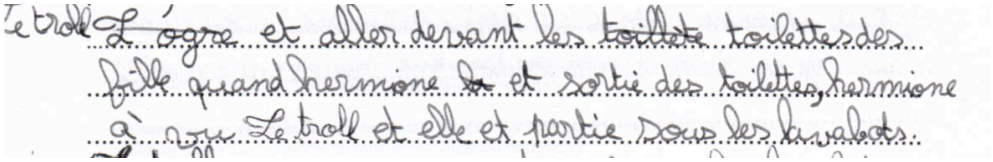
[2] A cet effet, nous travaillons également à décrire, selon la même méthodologie, les commentaires, corrections, indications insérés par les enseignants dans les copies de leurs élèves ; l'objectif à terme étant d'analyser précisément ce travail d'annotation des enseignants et de suggérer des propositions didactiques, complémentaires ou alternatives.

[3] Ces principes sont ceux qui prévalent dans les transcriptions présentées dans la revue *Genesis*, ainsi que dans la

Manuscrits modernes (ITEM) sont les suivantes :

- un mot ajouté est signalé entre chevrons : <ajout>
- pour le mot biffé, deux notations concurrentes subsistent :
- soit la biffure est notée de manière typographique : ~~biffure~~
- soit elle est signalée par un double crochet : [biffure]

Nous avons choisi de signaler les biffures entre crochets, de manière à ce qu'elles soient repérables automatiquement par un traitement de texte. Ainsi, l'extrait du brouillon suivant, d'un élève de fin d'école primaire[4] :



sera transcrit ainsi :

> <Le troll> [L'ogre] et aller devant les [toillète]toilettes des
> fille quand hermione [la] et sortie des toilettes, hermione
> à vu Le troll et elle et partie sous les lavabots.

Il existe théoriquement trois modèles de transcriptions : la transcription diplomatique, la transcription linéaire et la transcription chronologique. La transcription diplomatique est la moins interprétative des trois puisqu'elle se contente de restituer dans un espace graphique ce qui figure sur un autre espace graphique, en tentant de respecter l'ensemble des marques, telle une photographie. Au contraire, la transcription linéarisée repose sur l'interprétation en remettant sur un axe linéaire la succession des opérations d'écriture, selon une chronologie reconstruite. Dans l'exemple donné ici, on pourrait indiquer, à la fin de la première ligne, que l'élève a écrit « aller devant les toillète », biffé « toillète » et inscrit à sa place « toilettes ». Mais comment traiter la première opération visible, la biffure de « L'ogre » et l'insertion de « Le troll » ? Si l'on est certain, du fait de la position des GN sur la ligne, que « L'ogre » figurait avant et qu'il s'agit bien d'un remplacement, rien ne certifie du moment auquel il a eu lieu ; toute transcription linéaire serait, par conséquent, un choix interprétatif du transcripteur. Nous rejoignons J.-L. Lebrave (1990) quand il reproche à la transcription linéaire de ne proposer qu' « une interprétation univoque de la chronologie du manuscrit, qui devient contraignante si l'utilisateur n'a pas simultanément accès au document source ». Pour tenter de pallier ce problème, Lebrave a élaboré une méthode de transcription, dite *chronologique*, privilégiant la restitution des données temporelles et mettant au jour les différentes strates de l'écriture ; lui-même a présenté récemment ce travail en expliquant ses limites, mais aussi toute sa pertinence : « dès qu'on s'attaque à des manuscrits complexes, comme ceux de Flaubert, il devient rapidement évident qu'une reconstitution exhaustive de toutes les opérations, dans l'ordre où elles sont apparues, est impossible. […] En revanche, la reconstitution partielle d'opérations « locales » dans un fragment de manuscrit reste parfaitement possible » (Lebrave, 2009). C'est ce qu'a fait I. Fenoglio, par exemple, dans son étude sur les manuscrits autobiographiques d'Althusser où elle compare précisément la genèse d'extraits de deux textes du même auteur (Fenoglio, 2002), c'est également ce que nous pourrons

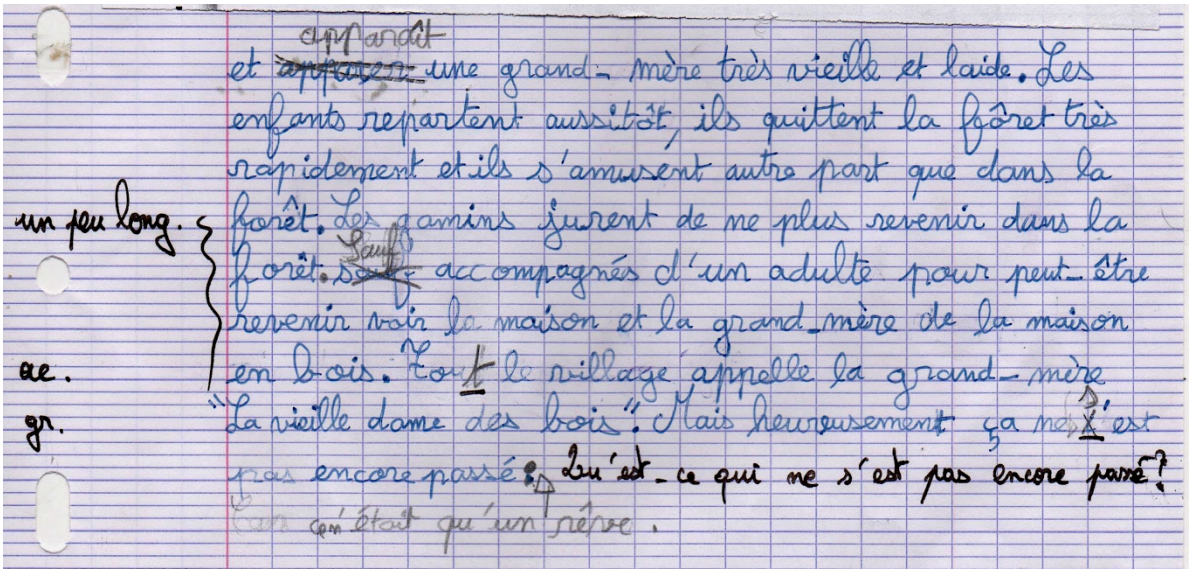
---

plupart des travaux des équipes de l'ITEM. Toutefois, la situation est loin d'être stabilisées et on ne peut que regretter l'absence de normalisation de la transcription diplomatique. Les indications données ici diffèrent des conventions suivies par d'autres institutions. A titre d'exemple, la Freie Universität de Berlin utilise la barre oblique pour indiquer un ajout de l'auteur et le double crochet pour indiquer un ajout de l'éditeur (http://www.geisteswissenschaften.fu-berlin.de/v/grammaire_generale/Principes_de_transcription/).

[4] Cours Moyen 2 en France, équivalent de la 5ème primaire.

faire pour aider à la lecture de brouillons courts et très surchargés ; mais de manière générale, nous suivons les linguistes généticiens pour privilégier la transcription diplomatique.

Voici un exemple de brouillon qui met en jeu deux scripteurs, l'élève, âgé de 10 ans (encre bleue et crayon gris) et l'enseignant (encre noire). Les deux instruments d'écriture utilisés par l'élève tracent une genèse en deux temps : l'écriture initiale (encre bleue) et la réécriture (crayon gris), parfois motivée par les remarques ou signes (soulignements) laissés par l'enseignant.




Voici présent la transcription correspondante (mis à part le commentaire en marge, traité plus bas). La lecture en est alourdie par les commentaires qui précèdent parfois les segments textuels ajoutés ou supprimés, et qui indiquent soit un scripteur différent, le professeur (code : P#) soit un moment d'écriture différé (code : T2#).

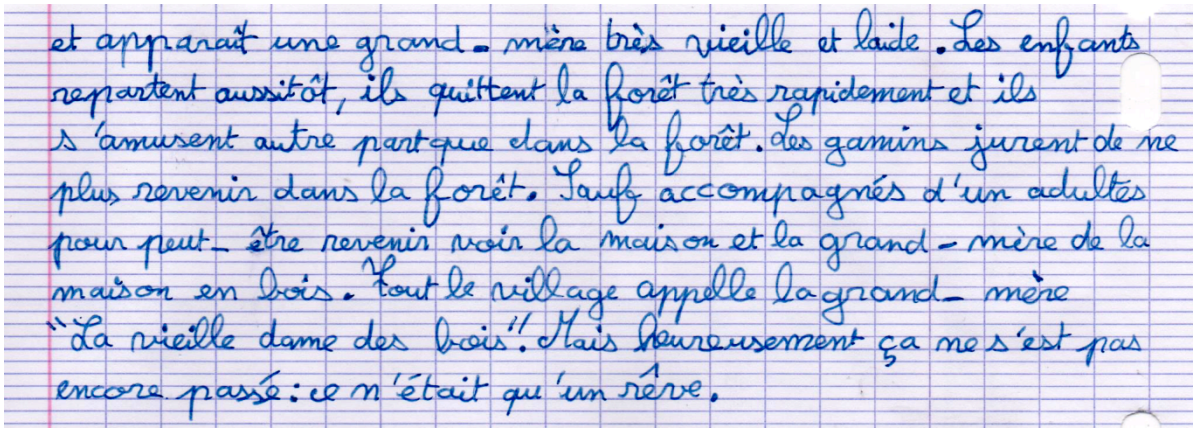
Transcription :

> et [T2#apparer]<T2#apparaît> une grand-mère très vieille et laide. Les
> enfants repartent aussitôt, ils quittent la forêt très
> rapidement et ils s'amusent autre part que dans la
> forêt. Les gamins jurent de ne plus revenir dans la
> forêt. [T2#sauf]<T2#Sauf> accompagnés d'un adulte pour peut-être
> revenir voir la maison et la grand-mère de la maison
> en bois. Tou<T2#t> le village appelle la grand-mère
> "La vieille dame des bois". Mais heureusement ça ne [c]<s>'est
> pas encore passé [T2#.]<T2#:> <P#Qu'est-ce qui ne s'est pas encore passé ?>
> [T2#Car] <T2#ce n'était qu'un rêve>.§

La première ligne de cette transcription pourrait se gloser de la manière suivante : à la suite de « et », on voit « apparer », qui a été supprimé dans une deuxième session d'écriture pour faire place à « apparait » ; ce segment est suivi de « une grand-mère très vieille et laide. Les ».

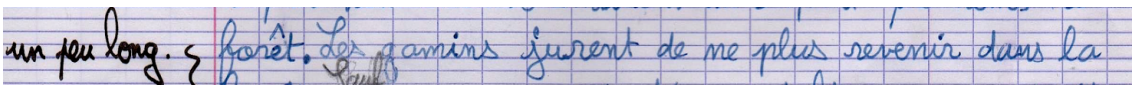
Voici comment gloser les deux dernières lignes : on lit le segment « pas encore passé », suivi d'un point qui, dans une deuxième session, fait place à deux-points ; à cet endroit le professeur a écrit « Qu'est-ce qui ne s'est pas encore passé ? ». A la ligne suivante l'élève a écrit, en deuxième session d'écriture, « Car », qui est supprimé, puis « ce n'était qu'un rêve ».

Ce brouillon a ensuite été recopié par l'élève pour conduire au texte final :


et apparaît une grand-mère très vieille et laide. Les enfants repartent aussitôt, ils quittent la forêt très rapidement et ils s'amusent autre part que dans la forêt. Les gamins jurent de ne plus revenir dans la forêt. Sauf accompagnés d'un adultes pour peut-être revenir voir la maison et la grand-mère de la maison en bois. Tout le village appelle la grand-mère "La vieille dame des bois". Mais heureusement ça ne s'est pas encore passé : ce n'était qu'un rêve.

Le mode de transcription adopté a l'avantage de permettre un accès direct, par le logiciel, aux différentes strates de l'écriture : scription initiale, éléments supprimés et ajoutés au cours de cette première session, éléments retravaillés au cours de sessions ultérieures, segments écrits par un autre scripteur, qui peut être identifié comme le professeur, etc. On pourrait donc, dans l'absolu, demander au logiciel de présenter l'ensemble des segments textuels qui ont été supprimés, et de les trier par session d'écriture ou par auteur. Il devrait également être possible - même si cela n'a pas encore été réalisé - de faire reconstituer automatiquement des états de texte en se basant sur la distinction entre opérations de première, deuxième, troisième session. Cette présentation ne pourrait certes présenter le détail de la chronologie, mais au moins les différents états du texte entre les sessions d'écriture.

Dans ce type de recherche, encore exploratoire, les niveaux technologique et linguistique se croisent et s'alimentent, chacun exigeant de l'autre, tour à tour, des adaptations des ses techniques de travail habituelles. Les questions de recherche posées par les brouillons et textes scolaires – ici la prise en compte de l'ensemble des événements d'écriture et en particulier des ratures – obligent les outils technologiques à s'adapter aux codages et aux données nouvelles. Mais la technologie contraint fortement, elle aussi, l'analyse linguistique : les règles de la transcription diplomatique, qui à l'origine « photographie » le document originel pour en rendre le maximum de caractéristiques, ont été légèrement modifiées, du point de vue du respect de la mise en espace. Par exemple, dans le brouillon reproduit plus haut, figure en marge un commentaire de l'enseignant, « un peu long », assortie d'un signe indiquant quelle zone du texte il commente. Ce type d'énoncé cumule deux caractéristiques dont le codage ne va pas de soi : il est commentatif, donc ne s'inscrit pas dans le texte lui-même, et il figure en marge. Pour la transcription, nous avons fait le choix suivant : le caractère commentatif est marqué par les signes { } mais l'emplacement en marge n'est pas figuré par la transcription, c'est-à-dire qu'on aura le même codage quel que soit l'endroit (marge, interligne etc.) où apparaît le commentaire. Ce dernier sera inscrit sur la ligne même :


un peu long. } forêt. Les gamins jurent de ne plus revenir dans la

{[P#un peu long]} forêt. Les gamins jurent de ne plus revenir dans la

Ce système permet au logiciel de localiser immédiatement l'ensemble des énoncés commentatifs. En revanche, on ne peut pas chercher automatiquement tous les énoncés en marge, *vs* par exemple les énoncés en interligne. Ce choix, fait au détriment du caractère photographique de la transcription, est un compromis entre la précision souhaitée et la commodité de traitement, évidemment nécessaire.

Dans les lignes qui suivent, nous présentons un autre objet de réflexion où s'est exercée de manière très étroite l'interaction entre questionnements linguistiques et contraintes technologiques : les annotations orthographiques des écrits.

## 1.2 Les écarts à la norme langagière et leurs conséquences sur l'annotation

Une des caractéristiques saillantes des écrits d'élèves est l'ampleur des écarts à la norme langagière, en particulier orthographique. Constitutifs de la langue écrite des énonciateurs, ces écarts ne sauraient être purement et simplement corrigés sans mettre à mal l'authenticité des discours analysés, donc celle des corpus constitués. Pourtant, ces écarts constituent un obstacle à l'étiquetage automatique puisque les formes non normées ne sont pas reconnues par le lemmatiseur. Il faut donc trouver un moyen pour que le logiciel lui associe, sans perdre la forme initialement produite par l'élève, la forme normée correspondante pour rendre possible la lemmatisation.

Avant d'exposer la solution technique que nous avons adoptée, il faut poser des limites de la normalisation. Voici la transcription du brouillon d'un élève de 10 ans (CM2, ou 5ème primaire) contenant de nombreux exemples d'écarts à la norme langagière :

> Ils [alla]<vont> voir! Et c'était un monstre tout [salle]<sale>
> il cassér tout, il manger tout. Ça veut dire que
> il y avait plus rien. Les enfant prévenir
> leur parents les parent [arrivère]<arrivent>. Le monstre est
> partit. ducout les parents ont crut que c'était les
> enfants. Les parents m'était leurs enfants dans leur
> chambre et les punicer.

On peut classer ces écarts de la manière suivante :

- sur- ou sous-segmentation : *ducout* (l.5), *m'était* (l.6) et d'élision avec apostrophe *que / il* (l.2-3) ;
- phonogrammiques n'affectant pas la valeur phonique : *salle* (l.1) ;
- morphophonogrammes grammaticaux : *cassér* (l.2), *manger* (l.2) ;
- morphogrammes catégoriels et parfois flexionnels : *enfant* (l.3), *leur* (l.4), *parent* (l.4), *partit* (l.5), *crut* (l.5), *m'était* (l.6) ;
- emploi des temps et modes verbaux : *prévenir* (l.3) mis vraisemblablement pour *préviennent* ; imparfaits : *m'était* (l.6), *punicer* (l.7), de la dernière phrase, là où on attendrait *a priori* des passés composés.[5]

Devant ce genre de production, la question se pose de savoir, pour chaque type d'erreur, s'il faut porter une annotation pour indiquer la forme normée. Nous répondons par l'affirmative pour les erreurs d'orthographe et de segmentation, de même que pour les erreurs morphologiques. Le choix a été fait d'établir un système d'annotation commun à ces deux types d'erreurs, il vaut aussi pour toute erreur, morpho- ou phonogrammique, altérant ou non la valeur phonique (par exemple : *pasion* mis pour *passion*), ce qui signifie que notre annotation n'offre pas pour l'instant la possibilité de distinguer entre les différentes catégories d'erreurs orthographiques et morphologiques (Catach, 1980a). Rien ne nous n'empêchera d'ajouter, par la suite, des couches d'annotation, le système actuel permettant de toute façon de retrouver les formes initiales.

Concernant les erreurs d'emploi des temps, elles posent des problèmes différents. Si nous proposons une annotation orthographique, c'est uniquement pour permettre à l'étiqueteur de fonctionner correctement, condition d'une analyse efficiente en lemmes et en catégories lexicales. Pour le reste des écarts à ce qui constituerait la norme, d'une part cette norme est parfois discutable (par exemple, la ponctuation dans les textes d'élèves est souvent lacunaire : comment la corriger ?), d'autre part il est important de pouvoir

[5] Le passé composé est suggéré en remplacement de l'imparfait car le texte est, dans les phrases précédentes, ancré au présent. Mais les deux modifications effectuées par le scripteur sur les temps verbaux (*alla* fait place à *vont*, *arrivère* à *arrivent*), marquent une oscillation du texte entre le présent et le passé. Et l'infinitif prévenir, dont on ne sait s'il est mis pour *préviennent* ou *prévinrent* (cette dernière forme n'étant sans doute pas connue de l'élève), maintient le flottement ; de sorte qu'il est difficile de se prononcer de façon certaine sur le temps verbal de la dernière phrase, dont on peut seulement considérer qu'il doit être non-sécant.

soumettre à l'analyse des textes qui soient les plus proches possible de la forme initialement produite par les élèves ; ne serait-ce que pour pouvoir étudier une partie de ces écarts – par exemple, pour les temps verbaux, un logiciel pourrait relever automatiquement l'ensemble des temps utilisés dans un texte, hors discours rapporté, et prédire une disjonction de la cohérence temporelle.

Tout cela considéré, il a donc été décidé de n'annoter, du point de vue des écarts à la norme, que les erreurs d'orthographe et de morphologie. L'annotation se fait de la manière suivante :

- le fichier annoté juxtapose à chaque forme erronée, la forme normée correspondante ;
- les balises < > encadrent chacune des formes ;
- le lien de substitution entre les deux formes balisées est marqué par _

Exemple : pour « il cassér tout », l'annotation, pour l'instant exclusivement manuelle, prend la forme suivante :

il <cassér>_<cassait> tout

La première forme encadrée de chevrons est la forme originelle ; elle est suivie de la forme normée. Avec cette solution technique on est assuré, pour chaque forme normée, de retrouver l'ensemble des graphies des élèves dans leur contexte d'origine.

## 3. La ponctuation blanche : un problème étendu dans les écrits d'élèves ?

Le terme de *ponctuation blanche* a été introduit par M. Favriaud (2000) dans son étude sur la poésie contemporaine pour désigner l'ensemble des espaces blancs qui organisent le texte poétique, en complémentarité à la ponctuation noire composée des signes de ponctuation et de divers marquages typographiques (le gras par exemple). Si le terme lui revient, M. Favriaud n'est pas le premier à s'être intéressé aux blancs de l'écrit : N. Catach (1980b) considérait déjà l'espace blanc comme « un signe de ponctuation en négatif », « le plus primitif et le plus essentiel de tous ». J. Anis (1983, puis Anis *et al*. 1988) en a fait un ponctuant à part entière en considérant l'ensemble des blancs – y compris le simple espacement entre les mots – comme concourant à la mise en espace qui est le propre de l'écrit.

Nous rassemblons sous cette dénomination de « ponctuation blanche » un ensemble de procédés qui relève de la ponctuation prise dans sa globalité (Pétillon, éd., 2004), mais aussi une série de phénomènes spécifiques qui apparaissent singulièrement dans les écrits d'élèves. Ainsi, les retours de ligne, les alinéas, les démarcations de paragraphes, de parties ou de chapitres... constituent les traces d'une organisation syntaxique et textuelle imposée par l'auteur – et parfois par l'éditeur – au(x) lecteur(s) ; traces qui, la plupart du temps, relèvent d'une volonté sciemment assumée par cet auteur. Mais qu'en est-il pour les élèves qui ne maitrisent pas encore ces procédés, et qui tentent d'organiser leurs textes en respectant des codes souvent peu enseignés, et contraints par les supports scolaires : la page du cahier (ou de la copie) avec ses marges et ses lignes verticales et horizontales ? Peut-on intégrer à l'analyse de cette ponctuation blanche des écarts révélant une acquisition en devenir, comme la segmentation en mots, qui n'apparait que dans les copies d'élèves[6] ?

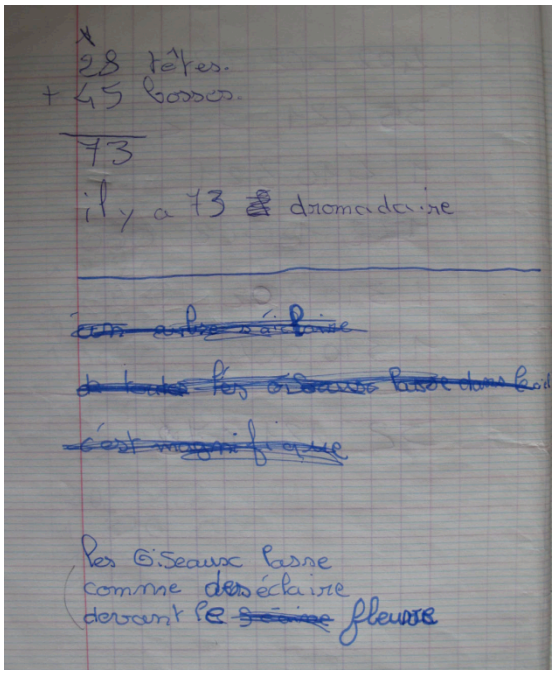
### 1.3 La segmentation en blocs de texte : ponctuation blanche et ponctuation noire

Les écrits des élèves, et singulièrement leurs brouillons, se caractérisent par un usage spécifique de l'espace graphique. Cet espace est contraint par le support d'écriture imposé par l'école, le plus souvent un cahier en primaire, au sein duquel apparaissent des zones délimités par des blancs. Ces zones sont

[6] Ou plus largement d'apprenants ne maitrisant pas encore les règles de distribution des mots, des phrases, des paragraphes et des textes en français.

repérables automatiquement puisque chaque retour à la ligne délibéré et chaque saut de ligne est marqué dans la transcription par le symbole du paragraphe (§).

Voici un exemple d'usage polysémiotique du blanc, tiré d'un cahier de brouillon de Cours Élémentaire 2 (3ème année de l'école primaire) :

| Manuscrit | Transcription |
|---|---|
|  | [1]<br>28 têtes§<br>+ 45 bosses§<br>____<br>73$<br>il y a 73 [ch] dromadaires§<br>_____<br><br>[un arbre s'aiclaire] §<br>§<br>[de toutes les oisaux passe dans le]ciel<br>§<br>[c'est magnifique] §<br>§<br>§<br>§<br>les Oiseaux Passe §<br>comme des éclaire §<br>devant [la]le [XXX] fleu[rs]ve § |

Cette page de cahier de brouillon comporte, comme c'est l'usage, des traces relevant de disciplines variées, ici les mathématiques et la poésie ; deux versions d'un même poème, repérable à des régularités lexicales (*éclaire/éclair*). Le découpage de la page en zones se fait, entre les deux disciplines, par un trait horizontal ; dans la partie « poésie », on remarque une mise en espace variante :

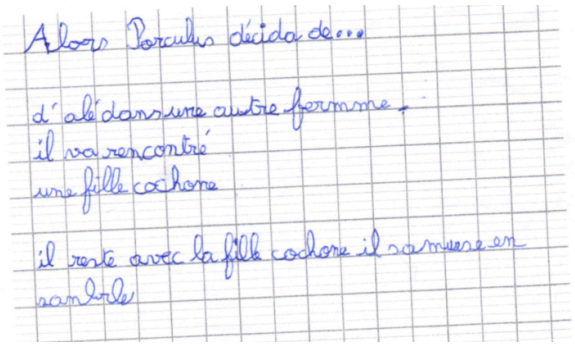
- la première version de poème, entièrement biffé, comportait des sauts de ligne systématiques ;

- plusieurs sauts de ligne le séparent de la deuxième version du poème.

Le blanc, éventuellement renforcé par un trait, paraît ici séparer différents blocs qui se distinguent soit du point de vue thématique (mathématiques / français) soit du point de vue de la logique scripturale : une première version d'un texte, écrite en sautant les lignes, sans doute en prévision de correction, puis entièrement biffée, est séparée par trois sauts de ligne de la seconde version. Cette distinction peut également renvoyer à un écart temporel : rien ne dit le temps écoulé, mais il peut être de plusieurs heures, voire d'un jour, entre l'écriture de la première version, sa biffure et l'écriture de la deuxième. Elle renvoie

de toute manière à une disjonction procédurale : j'écris / je biffe / je réécris, en des temps T1, T2 et T3. Ici, plus qu'un ponctuant de texte, le blanc est un ponctuant d'écrit, il trace des zones à l'intérieur d'un support, mais aussi en quelque sorte un ponctuant d'écriture : en séparant des blocs de texte, il délimite des tâches, et peut-être des périodes scripturales.

Une densité importante de blancs comme séparateurs de blocs d'écrits dans des cahiers de brouillon ne surprend pas, ce cahier étant précisément destiné à recevoir des traces diverses, souvent des bribes d'écrits plutôt que des écrits entiers. Plus surprenants sont les blancs, exprimés par des retours à la ligne et des sauts de ligne, dans le brouillon de récit suivant (il s'agit d'écrire la suite d'une histoire dont le début a été lu en classe, au Cours Élémentaire 1[7]) :

| Brouillon manuscrit de texte narratif | Transcription |
|---|---|
|  | Alors Porculus décida de… §<br>§<br>d'alé dans une autre fermme. §<br>il va rencontré §<br>une fille cochone §<br>§<br>il reste avec la fille cochone il samuse en samble § |



• Le premier saut de ligne marque un décalage énonciatif : la première ligne a été imposée par l'enseignante comme amorce du texte de l'élève et le saut de ligne semble indiquer une difficulté à s'inscrire « naturellement » dans cette continuité, difficulté qui peut être liée à la polyphonie ; ce blanc est renforcé par une autre ponctuation, les points de suspensions, que l'enseignante pensait certainement suppressibles mais que cet élève a maintenus, comme une frontière entre deux segments textuels dont les auteurs diffèrent. En observant l'ensemble des copies de cette classe sur la même tâche, on s'aperçoit que, si le maintient des points de suspension est rare, la mise en continuité reste problématique, comme le montre par exemple des débuts de texte tels que :

> Alors Porculus décida de alé dan
> une otre férme
>
> ou bien :
> allors porculus décida de …
> de… il va [r]chérché de la boue dan une
> grange

De plus, et à côté de ces résistances à la continuité linéaire qui se trouvent ici marquées verbalement, par l'absence d'élision d'une part et la répétition de la préposition *de* en début de ligne d'autre part, on trouve dans à peu près la moitié des copies un passage à la ligne entre la première ligne, imposée par l'enseignante, et la deuxième, écrite par l'élève ; ce passage à la ligne, qui n'est pas imposé par la dimension de la feuille, est un indice non verbal mais solide d'un « n'allant pas de soi » dans la tâche de

[7] Le texte lu en classe est *Porculus*, d'Arnold Lobel (Paris, École des Loisirs, 1971). C'est l'histoire d'un petit cochon qui adore se rouler dans la boue et qui est très déçu lorsqu'il découvre que la fermière, dans une frénésie de ménage, a tout nettoyé, y compris sa soue et sa mare. Cela va le conduire à quitter la ferme où il vit.

continuation de texte, qui est ici d'origine énonciative, en lien avec une interdiscursivité forcée dans laquelle les élèves peinent à s'inscrire.

• Le deuxième saut de ligne, entre les deux blocs que constitue le texte de l'élève, paraît lié à un changement d'épisode narratif : pour reprendre le schéma bien connu (Larivaille, 1974, puis Adam, 1992[8]), on peut penser que les lignes qui le précèdent constituent les péripéties du texte et celles qui le suivent, la situation finale. Le blanc délimite donc ici des unités thématiques. M. Fayol (1989) et B. Schneuwly (1984) ont constaté que la ponctuation, dans les premières années de l'écriture (CP-CE2), se manifeste de manière majoritaire « lors du passage d'une énonciation à une autre, [c'est-à-dire] lorsque l'auteur juxtapose des annonces de nouvelles » (Fayol 1989 : 23) et à la frontière d'épisodes du récit et lors de passages de la description à la narration par exemple, ce qui fait dire à J.-M. Passerault que « les signes de ponctuation constituent un marquage, en surface, des relations inter-événementielles » (1991 : 87). Notre extrait en constitue un cas exemplaire, si ce n'est que la ponctuation n'est pas noire – celle qu'étudiaient M. Fayol et B. Schneuwly (1987) –, mais blanche. Cette conformité avec des résultats obtenus à maintes reprises sur la ponctuation noire invite une nouvelle fois à faire entrer la blanche – passage à la ligne et saut de ligne – dans l'ensemble des marques de la ponctuation textuelle.

• Restent les passages à la ligne, évidemment délibérés, dans le bloc intermédiaire :

> d'alé dans une autre fermme. §
> il va rencontré §
> une fille cochone §

Le passage à la ligne vient renforcer en ligne 1 un point de fin de phrase. Ponctuation blanche et ponctuation noire se superposent donc, même si – c'est très courant à cet âge – la majuscule qui devrait suivre le point ne figure pas.

Plus étonnant est le passage à la ligne entre deux parties d'une même phrase, *il va rencontré* et *une fille cochone*[9]. M. Favriaud a montré, à partir de la poésie contemporaine, que les blancs de fin de ligne jouaient un rôle de démarcation de groupes syntagmatiques et de distinction entre thème et prédicat, la nature des constituants induisant des types de découpage. Dans son article synthétique de 2004, il donne des exemples de vers, extraits de poèmes de Jacottet sans ponctuation noire, où les blancs viennent ponctuer les différents groupes syntaxiques (GNS / GV par exemple) et les contraste avec l'exemple suivant, toujours chez Jacottet :

> Je marche
> dans un jardin de braises fraiches
> sous leur abri de feuilles
> un charbon ardent sur la bouche

Le blanc ne démarque pas, ici, le GNS du GV. M. Favriaud considère que « on voit, par le blanc, que le statut du groupe sujet n'est pas le même selon qu'il est formé d'un groupe nominal ou d'un indice de personne » (2004 : 19). L'auteur montre également que le blanc peut « rethématiser » des éléments qui relèvent syntaxiquement du rhème, et en rhématiser d'autres (*ibid*.).

Revenons à notre texte d'élève : on observe dans

> il va rencontré §
> une fille cochone §

une configuration similaire, pour la première ligne, au poème de Jacottet : le blanc regroupe GNS et verbe ; il rompt en revanche le GV, en séparant le verbe de son complément direct. La coalescence entre pronom et verbe paraît liée au caractère atone du pronom, qui le met dans la dépendance prosodique du

[8] Le schéma quinaire (en cinq points) de P. Larivaille : état initial / transformation (provocation + action + sanction) / état final, est repris par J.-M. Adam (1992) pour caractériser la séquence narrative type : situation initiale /{complication /actions / résolution} / situation finale.

[9] Souvenons-nous que le héros, Porculus, est un petit cochon (*cf*. note 6).

verbe qui le suit. Il nous semble que l'on peut également considérer que le blanc, ici, thématise le verbe et rhématise le COD, proposant un découpage thématique de la phrase en deux unités : la première centrée sur un personnage, la second sur l'autre. J. David (2008) a montré que les premières autographies des jeunes enfants (entre 5 et 6 ans) mêlaient une prise de conscience phonique à des impératifs thématiques, en particulier en ce qui concerne la segmentation : les enfants écrivent en un seul « mot » un groupe qui converge vers un signifié unique. On peut faire l'hypothèse que le même genre de procédure va s'observer, quelques années plus tard, dans l'écriture de textes. La segmentation en mots étant globalement acquise, c'est le découpage en phrases qui devient problématique à ce stade de l'apprentissage où la ponctuation, même si elle a été évoquée, n'a pas été enseignée systématiquement.

## 1.4 La segmentation en mots : un phénomène spécifique de ponctuation

Dans le corpus d'écrits d'élèves en cours de constitution, ces problèmes de segmentation en mots apparaissent massivement chez les scripteurs les plus jeunes, même s'ils perdurent chez des élèves plus âgés, notamment dans les zones d'instabilité de l'orthographe, et plus particulièrement de l'idéographie, parfois discutables, comme les noms composés. Il est ainsi fréquent de relever de telles erreurs de segmentation, comme chez Alexis (7,2 ans, CE1 ou 2ème primaire) :

> Lucile <jour>_<joue> [o>_<au> <plénobile>_<playmobiles>
> Juliette <mê>_<met> <con>_<son> <cêrtête>_<serre-tête>

ou les deux noms en position d'objet sont potentiellement segmentables (*plénobile* pour *playmobiles* ou *play-mobiles*), ou composés avec un trait d'union (*cêrtête* pour *serre-tête*), ou encore sans trait d'union, comme chez Silvia (7,6 ans, CE1) qui rédige un texte comportant une phrase proche de celle du précédent élève :

> Juliette a des <boucdoreilles>_<boucles d'oreille>

Tandis que cette Juliette (7,5 ans, CE1), quelques jours plus tôt, préfère elle aussi souder les noms composés :

> <É>_<Et> ils <petidéjene>_<petit-déjeunent>

On constate ainsi que les procédures d'encodage phonographique prennent souvent le pas sur la valeur sémiographique des mots (Jaffré, 2004). Les erreurs de segmentation sont ainsi révélatrices d'une focalisation sur une stratégie alphabétique qui n'intègre pas encore la réalité orthographique, même si celle-ci est, aujourd'hui comme hier, soumise à des règles discutables et sujettes à variation[10].

Au-delà de ces tolérances possibles ou envisagées des graphies du français, la maitrise du système écrit reste fondamentalement liée à l'identification des classes de mots, avec une analyse paradigmatique qui passe généralement par des procédures de *commutation*, ou de *remplacement* pour reprendre la catégorisation des opérations de réécriture. Ainsi, les scripteurs débutants de notre corpus posent des problèmes linguistiques souvent récurrents – notamment autour de la définition de cette unité « mot ». Les erreurs de segmentation peuvent être dès lors discutées[11] avec des arguments que la linguistique génétique éclaire d'un jour nouveau, car les erreurs de segmentation observées ne correspondent pas toujours à un lexème ou une unité lexicale (Petit, 1999), identifiable par des blancs graphiques. Ces écarts orthographiques nous renseignent sur les fonctionnements ou les dysfonctionnements de la langue, notamment dans ce rapport tâtonnant de l'oral à l'écrit, où la (non) délimitation blanche des mots ressurgit sur les conceptualisations et les débats scientifiques relatives à ces unités.

Dans le même domaine, nous relevons, dans les écrits de ces jeunes scripteurs, d'autres problèmes de segmentation, notamment dans l'emploi de certaines locutions plus ou moins figées, comme dans le récit

[10] Les Rectifications orthographiques de 1990 intègrent notamment cette tolérance à la soudure ou à la segmentation des mots composés.

[11] Voir la critique de la « Notion de mot », deuxième chapitre de l'ouvrage dirigé par M.-J. Béguelin en 2000.

de Loïc (7 ans, CE1) qui amalgame les éléments dissociés pour obtenir adverbe homogène :

> et <toutacou>_<tout à coup> un dragon <aparè>_<apparait>

Les soudures de mots sont ainsi très fréquentes dans des configurations qui associent des déterminants et de noms, comme dans les écrits d'Alexis (7,3 ans, CE1), qui agglutine plusieurs segments, entre autres avec « maifrair » :

> <aujour dui>_<aujoud'hui>, <maifrair>_<mes frères> <sinstal>_<s'installent> sur la table du salon

et un mois plus tard, à 7,4 ans, avec « mêparant », en changeant certes le codage phonogrammique :

> pendant les vacances, <mêparan>_<mes parents> <ramas>_<ramassent> les <gros>_<grosses> oranges

D'autres classes grammaticales sont également soudées, comme le nom avec l'adjectif postposé, tel la phrase de Louis (7,2 ans, CE1) qui, comme dans le cas des noms composés, inscrit le « chateaufor » :

> Il était une fois dans un <châteaufor>_<château fort>

et, dans un récit produit quelques jours plus tard, recourt au même procédé avec un nom précédé d'une préposition (+ le déterminant normativement amalgamés) « aules », puis avec une autre préposition (avec un déterminant absent) « aguofre » :

> avec <ser>_<ses> cafer>_<cafés> <aules>_<au lait>
> La pâte <aguofre>_<à gaufre>

Parfois ces procédés affectent des syntagmes entiers, comme chez Naty (7 ans, CE1) qui agglomère un syntagme entier :

> Pablo <Bouadeleau>_<boit de l'eau>

ou comme Johnson (7,2 ans, CE1) qui associe systématiquement les mots de ses syntagmes verbaux :

> je <merévele>_<me réveille> je <fématroileste>_<fais ma toilette> je <pronmonpetidéjené>_<prends mon petit-déjeuner>

Plus fréquemment encore, la segmentation défaillante est liée à des éléments apostrophés (pronom+verbe ; adverbe (de négation)+verbe ; déterminant+nom). Elle apparait de façon plus systématique et semble plus difficile à surmonter, car le recours à la procédure paradigmatique de remplacement est peu efficiente. C'est le cas des écrits des élèves suivants :

> Nowen (7,9 ans, CE1) : Soudain il <sai>_<s'est> <mi>_<mis> à cracher du feu ... je te dit que je <laime>_<l'aime>
>
> Paolo (7,3 ans, CE1) : Alors le chapeau jaune de <lomme>_<l'homme> <sanvole>_<s'envole>
>
> Silvia (7,6 ans CE1) : Mes parents <mappelle>_<m'appelle>
>
> Julie (6,11 ans, CE1) : il y a un stylo qui ne (ve>_<veut> <pa>_<pas> <souvrire>_<s'ouvrir>
>
> Alexandre (7,0 ans, CE1) : une agrafeuse qui <noz>_<n'ose> pas agrafer
>
> Julia (7,4 ans, CE1) : <Lotre>_<L'autre> dit « <ses>_<c'est> <chouète>_<chouette> »

L'analyse de ces phénomènes doit également prendre en compte d'autres réalisations graphiques, notamment ceux de sous- (ou d'hypo-) segmentation et de sur- (ou d'hyper-) segmentation. Nous pouvons en effet constater que les sous-segmentations restituées jusqu'ici varient selon la taille des segments affectés, en fonction des classes de mots « agglutinables » ; majoritairement les déterminants, les pronoms et les prépositions, précédant ou suivant les noms et les verbes. Ces amalgames montrent que la catégorisation grammaticale n'est pas évidente en termes de solutions graphiques des jeunes élèves. Elle recouvre une réalité linguistique que L. Hjelmslev, ([1968] 1971) avait en son temps énoncée dans la

distinction entre *plérèmes* et *cénèmes*[12] ; une distinction de toute évidence problématique que les scripteurs débutants rencontrent sans pouvoir toujours la surmonter.

Le tableau ne serait cependant pas complet si nous ne relevions pas des cas de sur-segmentation, certes plus marginaux mais également éclairants des procédures tâtonnantes des élèves[13]. En effet, certains élèves du même âge ont tendance à scinder les mots sur la base de leurs syllabes orales ; soit ils les traitent comme des noms composés, comme Sylvia (7,9 ans, CE1) :

> et ont fait un <bau-naum-de-neige>_<bonhomme de neige>

soit ils procèdent aléatoirement, semble-t-il, avec des solutions difficilement interprétables, comme avec Alex (7,5 ans, CE1) :

> et il fera <cho>_<chaud> et les <fers>_<fleurs> <repou ceron>_<repousseront>

On peut en effet penser que Sylvia tente de restituer des mots monosyllabiques (« bau » pour « beau » ; ce qui donnerai un « beau-(n)homme-de-neige ») ; alors qu'Alex ne suivrait aucune logique phonologique ou graphémique, lexicalement ou grammaticalement explicable.

## 4. Conclusion

Ces phénomènes de délimitation graphique ou de la ponctuation blanche des paragraphes, lignes et mots, à rapprocher de la difficulté des élèves de tous niveaux à gérer la ponctuation globale de leur texte, reste à étudier les procédés émergents, décalés et plus ou moins normés dans au moins deux perspectives : i) qualitativement pour envisager la diversité des erreurs de segmentation attestées, afin d'offrir des descriptions précises des composantes du système écrit qui semblent plus fragiles à acquérir par les élèves et tout autant délicates à enseigner ; ii) quantitativement pour appréhender leur nombre respectif, en les rapprochant des classes grammaticales, unités linguistiques et organisations discursives affectées, ainsi qu'aux niveaux scolaires impliqués.

Cette double analyse est l'une des ambitions de la recherche Ecriscol décrite, pour partie, dans cette contribution. Les écrits de notre corpus, rassemblés, organisés et mis en ligne, doivent ainsi offrir un ensemble de données exploitables, tant au plan linguistique pour modéliser le fonctionnement de cette ponctuation blanche, qu'aux plans psycholinguistique et didactique pour proposer des apprentissages conséquents dans un cadre développemental cohérent.

## Références bibliographiques


Adam, J.-M. (1992). *Les Textes : types et prototypes*. Paris : Nathan.

Anis, J. (1983). Pour une graphématique autonome. *Langue française*, *59*, 31-44.

Anis, J., Chiss, J.-L. & Puech, C. (1988). *L'Ecriture : théories et descriptions*. Bruxelles : De Boeck.

Crasson, A. & Fekete, J.D. (2007). Structuration des manuscrits : Du corpus à la région. Item [En ligne : http://www.item.ens.fr/index.php?id=173027].

Béguelin, M.-J. (éd.) (2000). *De la Phrase aux énoncés : grammaire scolaire et descriptions linguistiques*. Bruxelles : De Boeck & Duculot.


[12] Dans sa conception de la langue comme organisation où le contenu et l'expression ont chacun une forme et une substance, Hjelmslev ([1968] 1971) nomme *plérème* l'unité d'analyse de la forme du contenu, et *cénème* l'unité d'analyse de la forme de l'expression.

[13] Sur ces phénomènes d'hyper- ou d'hyposegmentation, peu d'études ont été menées, exceptées celles d'E. Ferreiro et C. Pontecorvo (1993) et celle de M. Ros-Dupont (1995).

Catach, N. (1980a). *L'Orthographe française. Traité historique et pratique.* Paris : Nathan.

Catach, N. (1980b). La ponctuation. *Langue française*, *45*, 16-27.

David, J. (2008). Les explications métagraphiques appliquées aux premières écritures enfantines. *Pratiques, 139-140*, 163-187.

Doquet, C. (2011). *L'Écriture débutante. Pratiques scripturales à l'école élémentaire*. Rennes : Presses universitaires de Rennes, coll. « Paideia ».

Favriaud, M. (2000). *La Ponctuation : la phrase – dans la poésie contemporaine à partir des œuvres de Du Bouchet, Jacottet, Stéphan*. Thèse de Doctorat sous la direction de Gérard Dessons, Université de Paris 8.

Favriaud, M. (2004). Quelques éléments d'une théorie de la ponctuation blanche – par la poésie contemporaine. *L'Information grammaticale*, *102*, 18-23.

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

Fayol, M. & Schneuwly, B. (1987). La mise en texte et ses problèmes. *In* J.-L. Chiss, J.-P. Laurent, J.-C. Meyer, H. Romian & B. Schneuwly (éds.), *Apprendre/enseigner à produire des textes écrits* (pp. 223-239). Bruxelles : De Boeck.

Fenoglio I. (2002). Une photo, deux textes, trois manuscrits. L'archivage linguistique d'un geste d'écriture identifiant. *Langages*, *147*, 56-69.

Ferreiro, E. & Pontecorvo, C. (1993). Le découpage graphique dans des récits écrits d'enfants entre 7 et 8 ans. Étude comparative espagnol-italien. *Études de linguistique appliquée*, *91*, 22-33.

Grésillon A. (1994). *Eléments de critique génétique : lire les manuscrits modernes*. Paris : Presses universitaires de France.

Hjelmslev, L. ([1968] 1971). *Prolégomènes à une théorie du langage*. Paris : Minuit.

Jaffré, J.-P. (2004). Peut-on parler de sémiographie optimale ? *LIDIL*, *30*, 11-26.

Larivaille, P. (1974). L'analyse (morpho)logique du récit. *Poétique*, *19*, 368-388.

Lebrave J.-L. (1990). Déchiffrer, transcrire, éditer la genèse. *In* A. Grésillon, J.-L. Lebrave & C. Viollet (eds.), *Proust à la lettre : les intermittences de l'écriture* (pp. 141-205). Tusson : Du Lérot.

Lebrave J.-L. (2009). Manuscrits de travail et linguistique de la production écrite. *Modèles linguistiques*, *59*, 13-21.

Masai, F. (1950). Principes et conventions de l'édition diplomatique. *Scriptorium*, *4*(2), 177-193.

Passerault, J.-M. (1991). La Ponctuation. Recherches en psychologie du langage. *Pratiques*, *70*, 85-103.

Pétillon, S. (éd.) (2004). *La Ponctuation*. *L'Information Grammaticale*, *102*.

Petit, G. (1999). La double hybridation de l'unité lexicale. *Linx*, *4*0, 137-158.

Ros-Dupont, M. (1995). La segmentation non normée de l'écrit de l'enfant de CE1 : erreur ou étape obligée de l'apprentissage. *Liaisons-HESO, 25*-26, 97-117.

Schneuwly, B. (1984). *Le Texte discursif écrit à l'école*. Thèse pour le Doctorat, Université de Genève.