

Rendre « visible » l'écriture scolaire. Choix de transcription et d'annotation pour une mise en corpus

Making school writing visible and legible. Transcribing and annotating a corpus of text

Claire Doquet - Université de Bordeaux, Lab-E3D

Claude Ponton - Université Grenoble Alpes, Lidilem

Résumé : Cet article expose les bases théoriques et méthodologiques du corpus E-CALM. Ce corpus a été construit pour répondre à la nécessité de documenter les compétences scripturales des élèves en France à différents niveaux de la scolarité. E-CALM constitue un réservoir de données textuelles exploitables par la recherche et par la formation des enseignants. Les choix de traitement des manuscrits recueillis ont permis de faire émerger les traces de l'écriture des élèves et de leurs interactions avec les enseignants corrigeant et commentant les copies. Explorer avec des outils informatiques un grand corpus d'écrits scolaires permet de confirmer des connaissances déjà acquises mais aussi de faire apparaître des éléments nouveaux, invisibles à l'observation de données de faibles tailles et qu'un grand corpus tel qu'E-CALM (plus d'un million de mots) permet de révéler. Une fois produite, l'analyse linguistique conduit à des avancées didactiques : en faisant émerger des éléments discriminants des textes, corrélés à des variables didactiques et sociologiques, elle permet de proposer des protocoles d'enseignement ajustés aux contextes d'apprentissage.

Mots-Clés : écrits non standard ; annotation des opérations d'écriture ; corpus scolaire et universitaire ; développement des compétences scripturales

Abstract: This article presents the theoretical and methodological foundations of the E-CALM corpus. This corpus was built in response to the need to document the writing skills of students in France at different levels of schooling. E-CALM is a reservoir of textual data that can be used for research and teacher training. The choices made in the processing of the manuscripts collected have made it possible to bring out the traces of pupils' writing and their interactions with the teachers who corrected and commented on the copies. The study of a large corpus of school writing with the help of IT tools allows us to confirm what we already know, but also to bring to light new elements that are invisible when observing small amounts of data, but which

can be revealed by a large corpus such as E-CALM (over one million words). Once produced, the linguistic analysis leads to didactic advances: by bringing out discriminative elements in the texts, correlated with didactic and sociological variables, it makes it possible to propose teaching protocols tailored to the learning context.

Keywords: non-standard writing; annotating writing operations; school and academic corpus; developing writing skills

Introduction

Le projet E-CALM soutenu par l'ANR s'est attaché à outiller l'exploration de l'écriture scolaire dans différentes dimensions :

- constituer un grand corpus d'écrits scolaires, qui comporte aujourd'hui plus d'un million de mots ;
- le traiter de manière à le rendre accessible à l'analyse de discours outillée ;
- réaliser les analyses dans trois secteurs : l'orthographe, la cohérence textuelle, les interventions des enseignants sur les copies ;
- élaborer des modèles didactiques visant à opérationnaliser certains résultats des analyses.

Ces dimensions sont présentes dans le titre même du projet, quand on en déplie l'acronyme : *Écritures scolaires et universitaires : Corpus, Analyses Linguistiques, Modélisations didactiques*.

Nous traitons ici des deux premiers volets du travail : la constitution et le traitement du corpus.

Les écrits scolaires sont une catégorie bien spécifique de l'archive écrite. À la suite de corpus déjà constitués et mis à disposition en ligne (*EMA*¹, *Scoledit*², *EcriScol*³, *Littératie Avancée*⁴), le projet E-CALM (*Écriture scolaire et universitaire : Corpus, Analyses Linguistiques, Modélisations didactiques*) a mis au point des procédures de recueil et de traitement des données écrites permettant aux outils actuels d'analyse de données textuelles d'explorer ces corpus particuliers. E-CALM a également permis la constitution d'un corpus original dédié aux problèmes de cohérence et de cohésion textuelle : *RésolCo*⁵.

¹ <https://www.ortolang.fr/market/corpora/ema-ecrits-scolaires-1/v3#!>

² <http://scoledit.org/scoledition/>

³ <http://syled.univ-paris3.fr/ecriscol/CORPUS-TEST/>

⁴ <https://lidilem.univ-grenoble-alpes.fr/ressources/corpus/litteracie-avancee>

⁵ <http://redac.univ-tlse2.fr/corpus/resolco.html>

La première finalité du projet É-CALM est l'analyse linguistique : explorer avec des outils informatiques un grand corpus d'écrits scolaires permet de confirmer des connaissances déjà acquises mais aussi de faire apparaître des éléments nouveaux, invisibles à l'observation de données de faible taille et qu'un grand corpus (plus d'un million de mots pour E-CALM) permet de révéler. Une fois produite, l'analyse linguistique permet des avancées didactiques : en faisant émerger des éléments discriminants des textes qui seront corrélés à des variables didactiques et sociologiques, elle permet de proposer des protocoles d'enseignement ajustés aux contextes d'apprentissage.

1. « Corpus manuscrits » : un oxymore ?

Dans sa conférence aux deuxièmes journées de linguistique de corpus en 2002, François Rastier définissait un corpus textuel comme « un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés : (i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications. » (Rastier, 2002). La même année, inaugurant la revue scientifique en ligne *Corpus*, Sylvie Mellet distinguait les corpus « homogènes et exhaustifs », sur lesquels s'appuient par exemple les monographies, des « corpus échantillonnés » où « l'enjeu n'est plus celui de l'exhaustivité, mais celui de la *représentativité* » (Mellet, 2002, §3 et 4). Rappelant l'adage *more data, better data*, elle indiquait que les corpus visant la représentativité se veulent souvent *de référence*, c'est-à-dire capables d'alimenter des recherches variées sur des données de même nature.

Un des enjeux de la recherche E-CALM est bien de constituer un corpus suffisamment important pour obtenir des résultats dont la validité excède ledit corpus. Pour autant, une réelle représentativité est illusoire tant les nombreuses variations contextuelles impactent les productions. Malgré ce que véhiculent certains manuels, l'étude des écrits scolaires montre la difficulté à constituer un échantillon représentatif des écrits d'élèves de 6 ans, de 7 ans, de 10 ans... Même si, de toute évidence, les compétences scripturales évoluent avec l'âge et le niveau scolaire, les sujets scripteurs et les contextes sont tellement différents que les productions varient de manière extrêmement sensible. Si certains traits de l'écriture peuvent se décrire sur un plan développemental (notamment ce qui relève de la mise en graphie), il n'est pas possible à l'heure actuelle de définir le texte type d'un élève de tel ou tel niveau scolaire.

1.1 *Un corpus de référence*

Conçu comme un corpus « de référence » de l'écriture scolaire, le corpus E-CALM se trouve donc d'emblée dans une tension entre aspiration à et illusion de la représentativité. À cette tension problématique, les chercheurs du projet ont répondu de deux manières différentes en complémentaires.

La première réponse a été de réunir des ensembles de données qui répondent à des critères divers reflétant une variété de pratiques :

- des données longitudinales (sous-corpus *Scoledit*), permettant d'étudier les mêmes scripteurs tout au long de l'école primaire, avec des tâches suscitées par la recherche et ressemblant à des tâches données de manière courante (par exemple : écrire à partir d'images) ;
- des données écologiques (sous-corpus *EcriScol*), donnant accès à des écrits dont l'ensemble des modalités de production et de traitement (notamment la correction des copies) sont décidées par les enseignants et reflètent donc le quotidien de leurs classes, avec un suivi des élèves sur une année scolaire (3 à 5 devoirs par an) ;
- des données plus expérimentales (sous-corpus *Résolco*), avec une tâche unique proposée de l'école primaire à l'université, tâche visant à susciter chez les scripteurs une réflexion sur la cohérence de leurs textes et non représentative des tâches données le plus fréquemment par les enseignants.⁶

Dans ce grand corpus, on peut observer des faits linguistiques supposément transversaux aux tâches, comme la gestion de l'orthographe, mais aussi des faits plus nettement dépendants des tâches, par exemple la cohésion textuelle qui est gérée différemment selon la consigne d'écriture. L'examen des corrections des enseignants, quand il y en a, donne des indications sur leurs priorités et en conséquence les contraintes qui pèsent sur l'écriture.

La deuxième réponse des chercheurs fut de renoncer à une réelle représentativité pour parler de significativité : tel résultat peut être considéré comme *significatif* en lien avec tel ou tel critère de sélection des données (le niveau scolaire, le contexte socio-économique de l'établissement, le type de tâche d'écriture...), sans pour autant qu'il faille considérer le corpus lui-même comme *représentatif* de l'écriture scolaire.

⁶ La tâche choisie, imaginée par Michel Charollles (Charolles, 1978), est décrite dans d'autres contributions de ce volume.

1.2 L'unicité du manuscrit vs la logique statistique

Mettre à disposition un corpus suppose idéalement de rendre accessibles toutes ses caractéristiques pour en faire des descripteurs et des critères de l'analyse. Le traitement des données d'É-CALM s'est heurté, dès l'abord, à la perte que constitue toujours la conversion de manuscrits en fichiers de données, à travers l'opération de transcription. Au-delà de la significativité de tel ou tel constat d'ordre linguistique ou sociologique, l'étude de manuscrits ne peut s'affranchir de considérer une composante idiosyncrasique de l'écriture : le geste graphique. Irréductiblement personnels, le tracé des lettres, leur épaisseur, leur taille, leur agencement, mais aussi la position des segments d'écriture sur la page, la « mise en page » de l'écrit forment un ensemble visuo-spatial que l'école façonne sans parvenir à l'homogénéiser. C'est pour rendre compte de cette dimension visuo-spatiale que les généticiens du texte⁷ ont adopté, le plus souvent, la transcription diplomatique déjà utilisée par la philologie. Cette solution n'est que partiellement satisfaisante mais elle préserve au moins la physionomie générale des pages et permet de faire des hypothèses génétiques en partant de la position des différents éléments sur la feuille.

Quand on veut, comme c'est le cas dans le projet É-CALM, que les écrits - et l'écriture - puissent être analysés informatiquement, les difficultés commencent. Le manuscrit est rétif à la mise en tableau. Un des traits de l'écriture manuscrite est précisément de résister à la schématisation. Elle ne se plie pas facilement aux méthodes d'analyse des données textuelles qui, comme leur nom l'indiquent, ont pour objet le texte et non l'écrit, et qui ignorent la substance conférée par le geste scriptural.

Dans son exploration de l'histoire de l'imprimé, William Mills Ivins (1953) fait de la reproductibilité un des apports les plus déterminants de cette technologie. Au-delà de ses aspects techniques et économiques, il s'intéresse à ses conséquences cognitives : jusqu'à l'invention de l'imprimé, seules des chaînes verbales pouvaient être considérées comme répétables à l'identique ; l'imprimé répète aussi l'icône. Cette propension à la répétition est étrangère au manuscrit qui frappe, au contraire, par son unicité ; un manuscrit porte la trace de la singularité de sa naissance. Même si, bien entendu, on peut le reproduire et le diffuser sous la forme d'un fac-similé, il échappe à la sérialisation parce qu'il intéresse comme trace d'un mouvement, d'une élaboration. L'œuvre terminée intéresse avant tout par le résultat qu'elle constitue ; elle appelle, en quelque sorte, sa reproduction. Il en va tout autrement du manuscrit,

⁷ Voir les travaux de l'Institut des Textes et Manuscrits, par exemple dans la revue *Genesis*.

qui n'a pas vocation à être publiable et que, de fait, on publie parfois sans que cette publication corresponde à un projet initial de son auteur. Document privé, préparatoire, lieu de tâtonnements et d'explorations, lieu de liberté aussi puisque comportant bien souvent des traces conjointes au texte, le manuscrit s'oppose à sa reproduction plutôt qu'il n'y conduit. C'est donc à une sorte d'entrave à la destination des objets manuscrits que sont les écrits scolaires que nous nous livrons ici : mettre à disposition ces écrits, les muséifier en quelque sorte, les inscrire, en tout cas, dans une certaine fixité, revient à les considérer comme (provisoirement) aboutis. C'est une gageure de rendre visibles sans les figer ces états intermédiaires, témoignages d'un état provisoire du travail qui conduira au texte final. C'est cette gageure qui a animé le travail de conception et d'élaboration de la base E-CALM. Elle présente différents états de ce *work in progress* que constituent toujours les écrits des élèves : *in progress* entre les différents états du texte, du brouillon à l'écrit final, traçant la genèse d'un texte ; *in progress* aussi entre les différents écrits produits au long de l'année, et même de plusieurs années, reflets de la genèse d'un scripteur.

2. Les écrits scolaires : des écrits d'apprenants, en situation particulière

Les écrits scolaires sont d'abord des écrits. Ceux que nous mettons à disposition sont manuscrits. En tant que tels, ils relèvent d'un gigantesque ensemble de documents organisés spatialement, dans le périmètre de feuillets (ici, cahier ou copies) éventuellement rassemblés en recueils. Cette matérialité va, pour une part, délimiter les investigations.

2.1 L'écrit, objet plurisémiotique

Un des traits importants pour une description linguistique des écrits est l'opposition, que l'on peut aussi penser comme une articulation, entre l'écrit et l'oral. Cette articulation procède à plusieurs niveaux :

- Sémiotique : tandis que l'oral fait appel, pour véhiculer les messages, à des traits auditifs (ton, de la voix, volume et rythme de l'émission...), l'écrit s'appuie sur des traits visuels (tracé des lettres, type de caractères, disposition sur la page des différents segments...)

Croquis, ratures, surcharges, effets de liste... on observe dans ce manuscrit, premier brouillon d'un devoir de français en 2nde (écriture inventée), une variété sémiotique et visuo-spatiale que l'on peut retrouver chez des écrivains, par exemple Paul Valéry dans le brouillon de son poème « Été » (fig. 2).



En quoi ces deux brouillons sont-ils comparables ? Jean Bellemin-Noël (1979) a livré de celui de Valéry (fig.2) une lecture psychanalytique et plurisémiotique où il analyse notamment les dessins très riches dont il fait une clef de lecture du texte. Le brouillon de Seconde (fig.1) comporte aussi des dessins, mais on a l'impression qu'il s'agit seulement de traces d'une activité destinée à passer le temps, à se désennuyer plutôt qu'à alimenter l'écriture. C'est peut-être en partie vrai, ce n'est sans doute pas totalement vrai. Même si l'on ne peut identifier dans le

brouillon scolaire des dessins renvoyant directement au texte, il est vraisemblable que les petits croquis présents en marge haute de la feuille correspondent à des moments de recherche où la scriptrice a besoin d'utiliser son stylo même si elle n'inscrit pas d'éléments qui paraissent directement utile à l'écriture.⁸ La trace renverrait, de manière tautologique, au traçage, au tracé, plutôt qu'à un référent externe. L'inscription, pendant l'écriture, dans un autre système sémiotique, ici le dessin, mobilise aussi la psychologie développementale : dans ses réflexions sur la « préhistoire de l'écriture », Lev Vygotski (2014) range le dessin parmi les activités par lesquelles les enfants construisent des significations, et insiste sur l'importance de cette création ; les significations, au lieu d'être simplement vécues, sont créées par le sujet qui, explique Michel Brossard, « agit non plus en réponse à la situation extérieure mais en fonction de significations qu'il crée lui-même et attribue aux choses » (Brossard, 2013 : 26). Cette prise de conscience de la capacité créatrice est aussi une condition du passage à l'écrit ; le dessin, continue Brossard, « nous met sur la route de l'écriture » (*ibid.* : 27).

Ces propos décrivent le développement psychologique de l'enfant ; le dessin, comme le jeu ou la parole orale, constitue pour Vygotski un symbolisme de premier degré, tandis que l'écriture est un symbolisme de second degré. Dans notre copie de Seconde, la juxtaposition de dessins et de textes ne peut être interprétée strictement de cette manière. Toutefois, on peut penser que le dessin a une fonction programmatique, telle qu'on la retrouve chez Zola par exemple (Lumbroso & Mitterrand, 2002), parce que « le dessin est accompagné de discours » (Brossard, 2013 : 27) et que ce discours va se retrouver, au moins en partie, dans l'écrit associé au dessin.

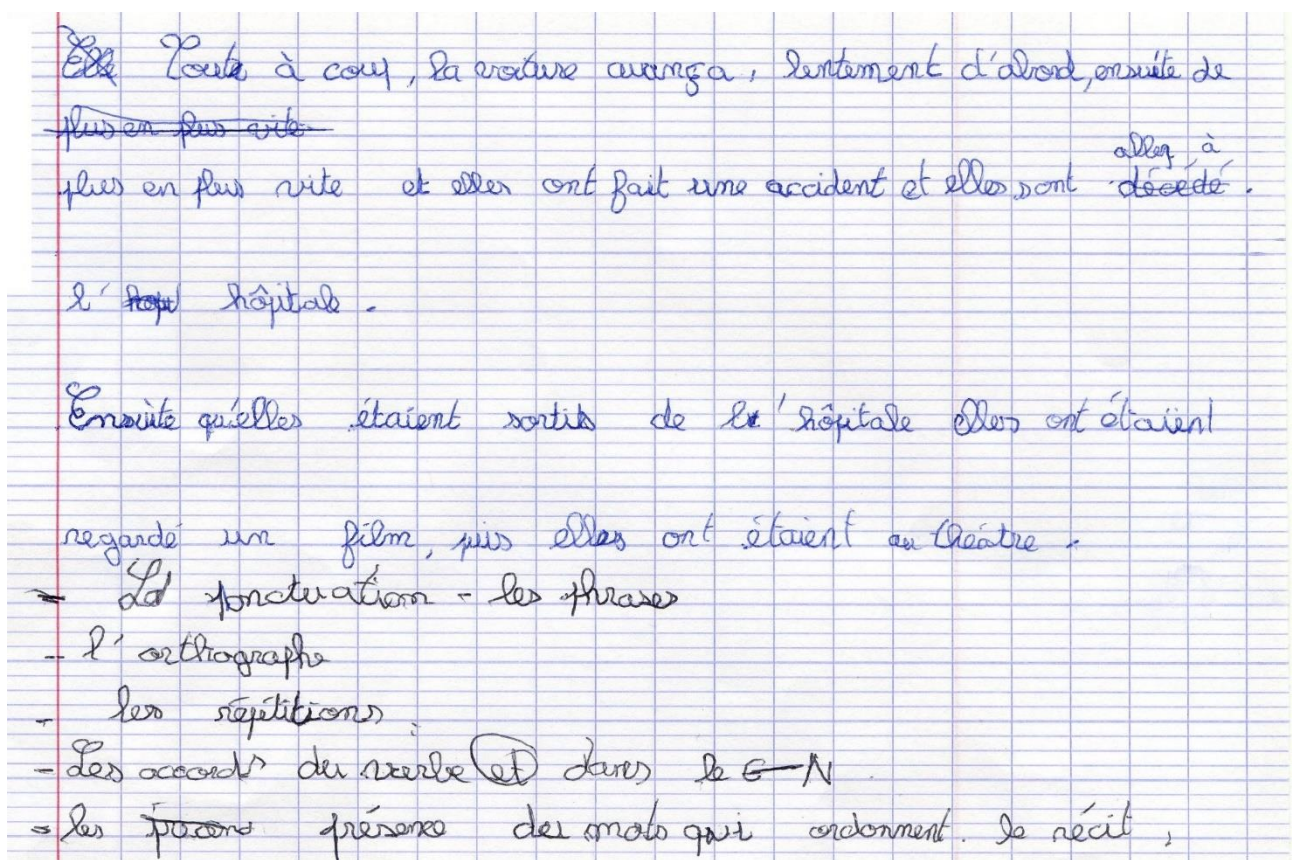
2.2 La substance graphique

Cette articulation entre systèmes sémiotiques est toujours à l'œuvre à l'écrit où, même si ne figurent pas sur la page des dessins à proprement parler, la substance est d'ordre graphique. Jean-Louis Lebrave, un des initiateurs en France de la génétique textuelle, évoquait dès 1999 la « puissance expressive de la substance graphique des manuscrits » (Lebrave, 1999 : 371). De fait, un manuscrit présente une variété sémiotique multimodale : sémiotique verbale, bien sûr ; sémiotique spatiale également ; sémiotique chromatique parfois, mais aussi iconique. C'est tout cela que Lebrave désigne sous le terme de « substance graphique » : l'ensemble des

⁸ De tels comportements s'observent aussi sur traitement de texte, dans des élèves, cherchant l'inspiration, manipulent frénétiquement les touches « retour » et « suppression » pour créer puis éliminer des lignes, ou bien déposent des signes divers (traits horizontaux, barre oblique, etc.) qu'ils suppriment illico (Doquet, 2011).

éléments porteurs de signification pouvant être isolés et dont l'interaction forme la « substance » du manuscrit. Cette substance⁹ devrait être prise en compte dans toute analyse. C'est ce que font les généticiens du texte, lorsqu'ils travaillent sur des manuscrits d'écrivains : noter le plus d'informations possibles sur l'aspect du document, l'état du papier, le ou les instrument(s) d'écriture, l'orientation des lignes, la nervosité du trait... tous ces éléments contribuent à l'impression donnée par le document, à côté de la substance verbale.

Que dire d'un écrit scolaire, où le jeu des couleurs et des formes est tout aussi riche ? En voici un exemple (fig. 3) :



Dans ce brouillon, on repère d'abord deux couleurs d'encre. Sémiotique chromatique, donc, renvoyant à la procédure de composition de l'écrit : deux instruments d'écriture, donc sans doute deux étapes.

⁹ Le terme renvoie à l'articulation saussurienne entre forme et substance, que Boe repose, rapidement mais clairement, comme « relations entre le système et sa réalisation » (Boe, 1997 : 8). En continuité avec Mahrer qui assimile la substance de l'expression écrite au signal graphique (Mahrer 2014 : 31), Doquet et *al.* (2022) parlent de « signaux graphiques et acoustiques » pour caractériser la substance des signes.

Pour faire des hypothèses sur l'ordre d'écriture des segments, on va se fier à l'organisation des énoncés sur la page. Sémiotique spatiale. En règle générale, l'écriture commence en haut à gauche ; il est donc probable que les segments en bleu ont été écrits avant ceux en noir. La connaissance de la situation de production de cet écrit confirme cette hypothèse : l'enseignante a relevé les brouillons écrits en bleu, elle les a lus sans les corriger, les a rendus aux élèves et leur a demandé de noter en noir des éléments sur lesquels ils devraient exercer une vigilance particulière lors de la poursuite de l'écriture.

3. Choix de transcription

Ceci posé, la question est de savoir comment présenter ces écrits, que retenir de l'ensemble des informations qu'ils nous donnent pour permettre aux usagers de la plateforme E-CALM d'en explorer au mieux la richesse. Le premier des choix opérés par l'équipe est de regarder ces écrits pour ce qu'ils sont, des étapes d'un *work in progress*, et de donner la première place aux traces de cette progression : retours dans le texte, ratures, ajouts d'éléments... autant de signes permettant de reconstruire, comme le font d'autres chercheurs sur les manuscrits littéraires, la procédure d'écriture des textes d'élèves. Cela signifie, en premier lieu : ne pas ignorer les événements d'écriture visibles sur la page mais au contraire les prendre en compte, en les transcrivant mais aussi en les annotant pour permettre aux outils d'exploration de les retrouver et de les traiter.

Pour satisfaire à ce choix prioritaire, nous avons utilisé les techniques explorées avant nous par la génétique textuelle, qui elle-même les hérite, pour partie, de la philologie. Il s'agit d'abord de trouver une représentation des manuscrits qui permette d'en donner le maximum de détails, en préservant les différentes sémiotiques repérées (spatiale, chromatique, iconique, verbale). Pour ce faire, nous choisissons dans un premier temps la transcription diplomatique (cf. *Supra*, 1.b), qui vise à restituer les éléments d'un manuscrit en respectant précisément les niveaux sémiotiques. La transcription diplomatique « photographie le document en rapportant, avec les outils qui le permettent, malgré leurs limites, tous les événements du manuscrit » (Crasson & Fedeke, 2004). Concrètement, le transcritteur transcrit tous les éléments du manuscrit, en respectant sa topographie, la couleur de l'encre, l'ensemble des ratures, et en indiquant s'il y a lieu des précisions sur d'autres aspects difficiles à transcrire (par exemple, un trait très appuyé, un pâti d'encre, etc.).

Dans cette logique, voici comment pourrait être transcrit (fig. 4) le brouillon de fig. 3 :

Figure 4 : transcription du brouillon de fig.3, élève de Cours Moyen 2

Elle Toute à coup, la voiture avança, lentement d'abord, ensuite de
~~plus en plus vite~~ allez à
plus en plus vite et elles ont fait une accident et elles sont ~~décédé~~.

l'h~~osp~~ hôpitale.

Ensuite qu'elles étaient sortis de l'hôpitale elles ont étaient

Regardé un film, puis elles ont étaient au théâtre.

La ponctuation - les phrases

l'orthographe

les répétitions

Cette transcription, bien que réputée la plus fidèle possible (Crasson et Fedeke parlent de « photographie » du manuscrit), constitue une perte par rapport à l'appréhension directe du manuscrit. Elle met en évidence les différents espaces de la page, notamment grâce à la couleur de l'encre. Elle tient compte des sauts de ligne opérés par le scripteur. En revanche, elle ne différencie pas certains éléments, comme les différents types de rature : croix sur le mot à supprimer (sur le premier mot, *Elle*), traits horizontaux répétés (sur *plus en plus vite*, par exemple). Et bien entendu, les caractères imprimés n'ont pas la même valeur que le manuscrit : la lisibilité gagnée par l'usage de l'imprimé crée une distance entre le geste d'écriture et le lecteur qui l'observe.

Cette manière de présenter les données nous a paru un compromis acceptable entre la volonté de conserver le maximum d'indices de la substance du manuscrit et la nécessité de rendre lisible, et exploitable informatiquement, l'intégralité du corpus. C'est finalement en tension entre lecture humaine et lecture informatique que se sont opérés les choix successifs. Nous allons à présent présenter quelques-uns de ces choix qui permettent de tirer parti des données ainsi rendues disponibles.

3.1 Choix techniques

L'un des objectifs du projet E-Calm est la mise à disposition auprès des communautés scientifiques et du monde l'éducation de la ressource constituée via des plateformes du type

Ortolang¹⁰ ou Nakala¹¹. Cette visée de large diffusion implique la création d'une ressource dans un format dit interopérable. Pour des données de nature textuelle, le choix du XML selon les recommandations de la TEI-P5¹² s'impose. Il garantit la cohérence, la praticabilité et la compatibilité avec un large éventail d'outils de traitement corpus et facilite l'échange de données (Burnard, 2007). Si les corpus en TEI recouvrent de nombreux domaines (corpus d'interactions sur les réseaux sociaux, corpus de textes scientifiques ou journalistiques, brouillons ou manuscrits d'auteurs, etc), il n'existait aucune description concernant les corpus scolaires. La proposition d'un guide de transcription et d'une DTD XML-TEI pour ce type de corpus est donc une avancée importante issue du projet E-Calm.

Si la TEI propose un outillage assez complet pour décrire les éléments purement textuels, notre proposition permet de prendre en compte les spécificités des corpus scolaires à savoir, d'une part, les opérations de réécriture réalisées dans des temporalités différentes et par deux mains possibles (élève ou enseignant) et, d'autre part, des commentaires enseignants portant sur des passages spécifiques ou sur la copie dans son ensemble.

Le texte

Les choix de transcription sur le contenu textuel découlent de deux objectifs.

Le premier est la possibilité de mener des analyses linguistiques sur les compétences en littéracie des élèves ce qui implique de conserver les graphies de l'élève telles qu'elles apparaissent dans le texte sans aucune forme de normalisation (respect de la casse, de l'accentuation, de la ponctuation, de l'orthographe, etc.).

Le deuxième objectif est de permettre un retour simple au fac-similé afin de visualiser l'ensemble des caractéristiques de la copie. Pour faciliter cette lecture en vis-à-vis du fac-similé et de la transcription, la mise en page (retours à la ligne dus à la marge, césures, etc.) sont transcrits tels quels. De plus, pour faciliter les traitements outillés, les éléments de structures apparents du texte comme les paragraphes et les titres sont également décrits

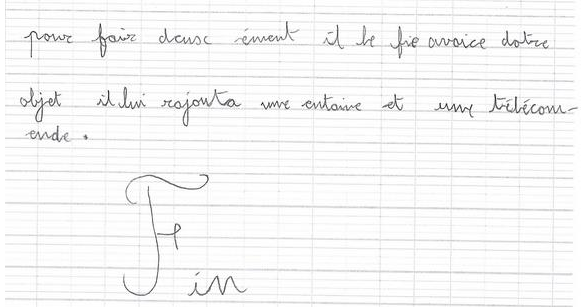
Par exemple, dans l'extrait d'une copie de CE2 suivant (fig.5), on trouve deux paragraphes (la fin de l'un et le paragraphe contenant « *Fin* ») avec deux retours à la ligne dans le premier. Le deuxième retour comporte en plus une césure sur le mot « *télécomende* »

¹⁰ <https://www.ortolang.fr/>

¹¹ <https://www.nakala.fr/>

¹² <https://tei-c.org/guidelines/p5/>

Figure 5 : un manuscrit et sa transcription

 EC-CE2-2016-61-D1-S2028-V1 (élève 2818, CE2, corpus Scoledit)	Transcription pour fair deux ément il le fie avaiice dotre <lb/> objet il lui rajouta une entaine et une telecom<lb rend="hyphen"/>ende.</p> <p>Fin</p>
---	--

Dans la transcription (fig.5), le texte de l'élève est conservé sans modification. La balise <lb/> désigne un retour à la ligne du aux marges ; l'attribut rend marque le retour à la ligne avec une césure. Enfin, les balises <p> et </p> encadrent chaque paragraphe du texte.

Les temps d'écriture

Les copies du corpus E-Calm ne sont pas toutes composées d'un seul et unique premier jet comme c'est le cas dans les corpus Scoledit ou Resolco. Elles résultent d'aller-retours entre élève et enseignant et proposent des états de texte divers allant du brouillon à la copie finale. L'écriture peut alors être découpée séquentiellement en unités temporelles, équivalant à ce que la génétique textuelle appelle « campagnes d'écriture ». Nous avons adapté la notion à la situation spécifique de l'écriture scolaire où la personne qui intervient sur la copie peut être l'élève auteur, mais aussi l'enseignant. Le terme « campagne d'écriture » ne paraissant pas adapté à la phase de correction par l'enseignant, nous parlons de « temps d'écriture ». Un temps d'écriture est caractérisé par le sujet qui écrit (élève ou enseignant, appelés *mains* dans le cadre de la TEI) et par sa séquentialité, représentés à l'aide des attributs TEI suivants :

- ⇒ *resp* : décrit la nature de la personne responsable de l'intervention (ie. E pour élève et P pour professeur).
- ⇒ *seq* : indique le temps de l'intervention (ie. T1 pour le premier jet, T2, T3...)

Les opérations d'écriture

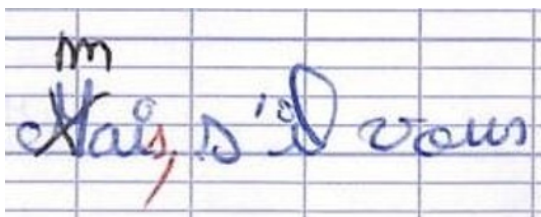
Dans la tradition de la philologie et de la génétique textuelle, nous codons les opérations d'écriture de base (*ie.* suppressions, insertions, remplacements et déplacements) comme des modifications du texte réalisées dans un temps d'écriture donné.

En TEI, ces modifications sont codées à l'aide de la balise *mod* puis spécifiées avec les balises *add* et *del*. Les remplacements et les déplacements sont vus comme une suppression et une insertion.

Dans l'exemple suivant (fig.6), extrait d'une copie de CM2, on trouve trois temps d'écriture caractérisés par des couleurs d'encre différentes avec un remplacement et un ajout :

- Temps 1 : en bleu, premier jet de l'élève
 - Transcription : **Mai s'il vous**
- Temps 2 : en noir, révisions de l'élève
 - Transcription du remplacement de « M » par « m » :
<mod type="subst" resp="E" seq="T2">M<add>m</add></mod>
- Temps 3 : en rouge, intervention de l'enseignant
 - Transcription de l'ajout de « s, » :
<mod type="add" resp="P" seq="T3"><add>s,</add></mod>

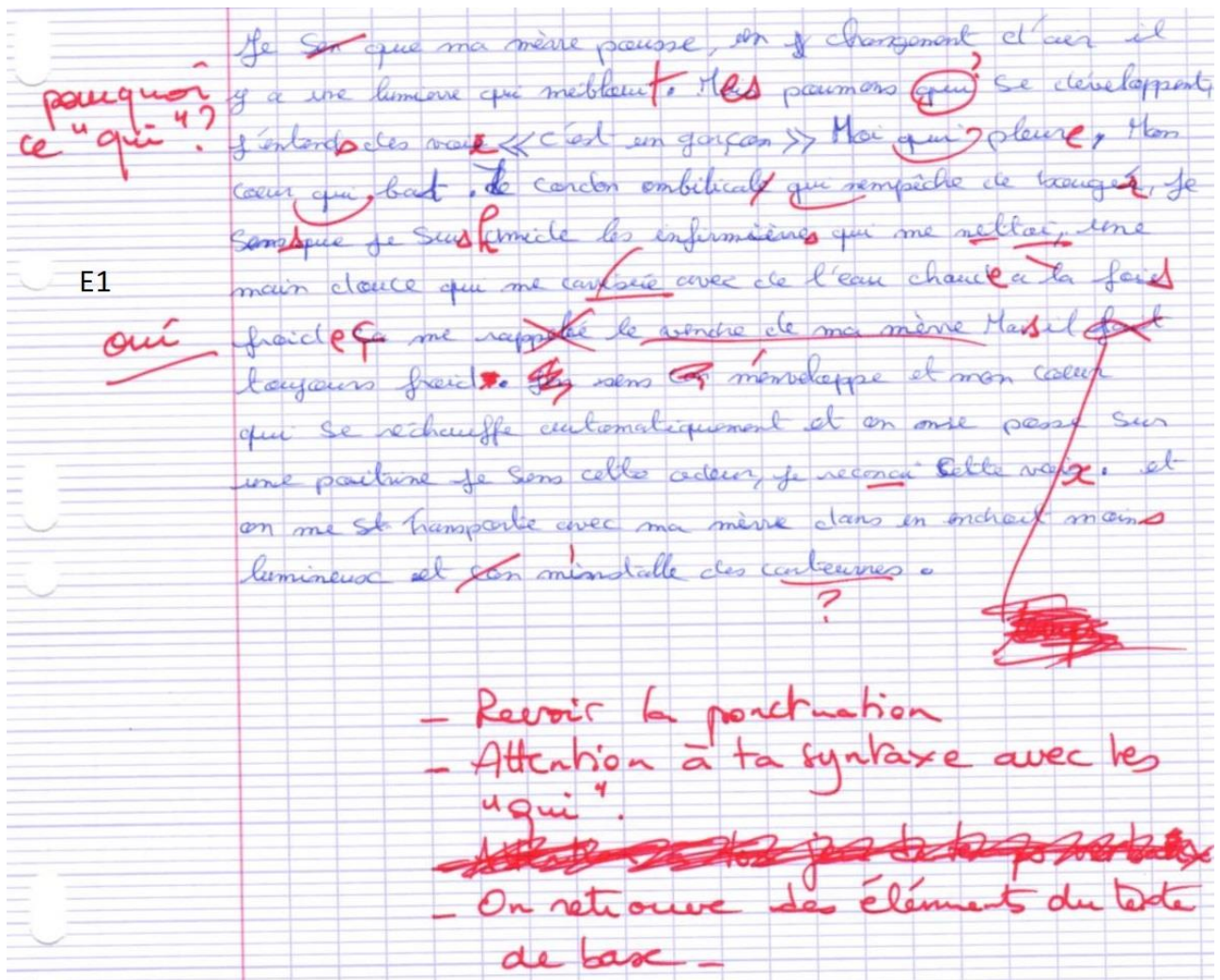
Figure 6 : manuscrit et transcription pour la partie « les opérations d'écriture »



Les commentaires enseignants

Dans notre corpus, les commentaires des enseignants sont de deux types : les commentaires spécifiques à un passage du texte et les commentaires généraux portant sur l'ensemble de la copie.

Figure 7 : manuscrit d'élève avec commentaires de l'enseignant



Outre des interventions directes de l'enseignant sur le texte, la copie de 3^{ème} ci-dessus (fig.7) montre ces deux types de commentaires :

- Commentaires spécifiques : « pourquoi ce "qui" ? » qui porte sur le mot qui entouré par l'enseignant et « oui » qui porte sur le passage souligné « le ventre de ma mère »
- Commentaire général en fin de copie avec 3 alinéas

Les commentaires généraux sont simplement transcrits à l'aide des balises <metamark> de la TEI et placés avant (<front>) ou après (<back>) le texte selon leur position.

Dans le cas de l'exemple ci-dessus, la transcription donne :

<metamark resp="P" seq="T2">

1. - Revoir la ponctuation<lb/>
2. - Attention à la syntaxe avec les<lb/> "qui ". <lb/>

3. `<mod type="del" resp="P" seq="T2"><del>- Attention <gap/> jeu de temps verbaux</del></mod>`

- On retrouve des éléments du texte<lb/> de base.

</metamark>

Les commentaires spécifiques sont transcrits en deux parties :

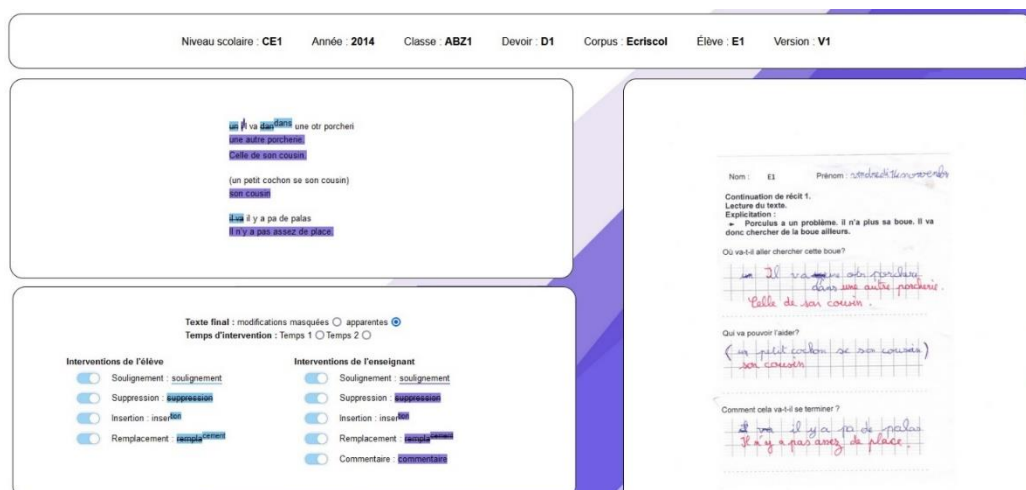
1. Marquage dans le texte de la zone sur laquelle porte le commentaire à l'aide des balises `<span/>` (début de la zone) et `<anchor/>` (fin de la zone).
2. Transcription du commentaire après le texte avec un lien sur la zone marquée avec une éventuelle précision sur la place de ce commentaire dans la copie

Dans l'exemple ci-dessus, le commentaire « pourquoi ce "qui" ? » est transcrit de la manière suivante :

- Marquage de la zone ciblée :
4. `<span xml:id="c1" to="#endC1"/>qui<anchor xml:id="endC1"/>`
- Commentaire spécifique :
`<note place="margin-right" target="#c1">Pourquoi ce "qui" ?</note>`

4. Une proposition pour rendre lisible le corpus

L'ensemble des conventions exposées ci-dessus, forgées par les participants au projet É-CALM, permettent bien entendu de rechercher, à l'aide d'outils adaptés, les différents types d'événements d'écriture. C'est aussi grâce à ces conventions que l'on parvient à donner une représentation lisible des écrits scolaires, sur la plateforme dédiée (fig.8) :



Ce mode de présentation permet de différencier :

- en bleu, les corrections de l'élève, ici réalisées au moment même de l'écriture et donc affectées au « temps 1 »
- en violet, les interventions de l'enseignant, réalisées en « temps 2 ».

En décochant les cases en bas à gauche, on peut sélectionner la temporalité mais aussi les opérations scripturales, pour faire apparaître par exemple les ajouts réalisés en temps 1, les commentaires faits en temps 2, etc. Cette présentation permet de déplier la chronologie de l'écriture pour en isoler les différentes strates et ainsi mieux en comprendre la logique. Sur cette copie, on dissocie : les événements d'écriture survenus en temps 1 (figure 9) de ceux qui sont survenus en temps 2 (figure 10).

Figure 9

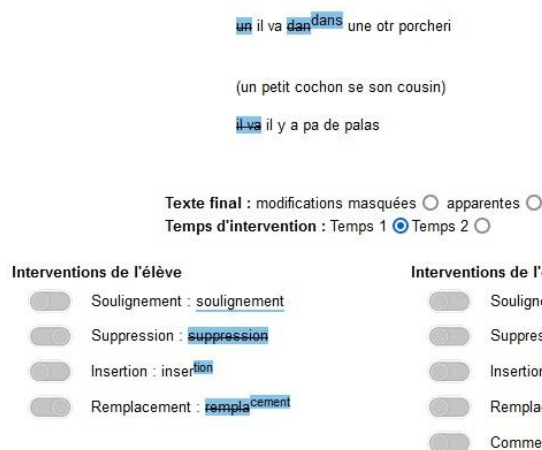
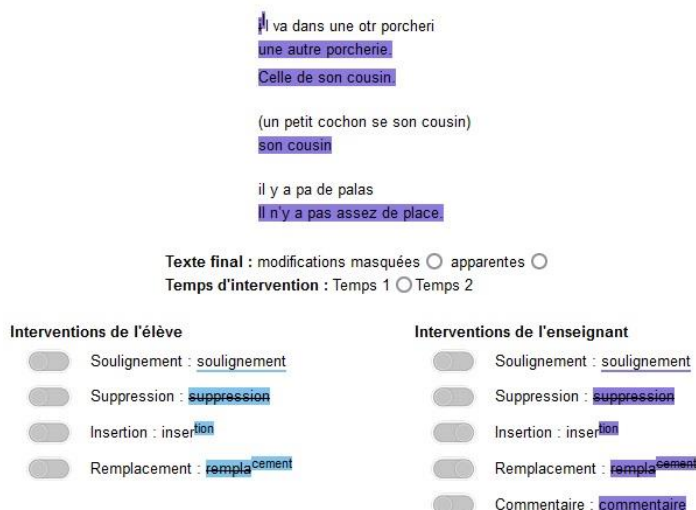


Figure 10



Au-delà de la lisibilité, dont la plus-value est évidente dès lors qu'on a affaire à des copies complexes réalisées en une série d'étapes et que les nombreuses ratures et reprises finissent parfois par rendre illisibles, une telle présentation est adaptée à la volonté de rendre visible l'écriture elle-même, dans une perspective génétique. Elle peut aussi être un outil intéressant pour l'enseignement, en permettant de revenir à un état antérieur du texte qu'il peut s'agir de faire retravailler, en gardant sous la main les opérations effectivement réalisées par son auteur qui pourront ensuite être questionnées. Le travail de codage des copies et les choix de transcription et d'annotation permettent donc de travailler sur l'écriture elle-même, alors que c'est souvent le texte, et derrière lui la correction linguistique, qui sont interrogés à l'école. Une

des innovations de la recherche É-CALM est d'avoir pu outiller informatiquement les préconisations des chercheurs en didactique de l'écriture qui s'intéressent, depuis une trentaine d'année, aux écrits scolaires comme texte en devenir, et aux élèves comme scripteurs en construction.

5. Perspectives

La plateforme de visualisation E-CALM est disponible en ligne dans une version intermédiaire (<https://e-calm.huma-num.fr>). Elle continue d'être travaillée dans deux directions complémentaires :

- L'augmentation du nombre de copies disponibles, pour permettre aux utilisateurs de naviguer entre les niveaux de classe et les contextes des établissements.
- La facilitation des recherches critériées, pour permettre de sélectionner des écrits répondant aux questionnements des utilisateurs : recherche en plein texte autour d'un mot par exemple, mais aussi tri des textes selon le taux d'opérations de réécriture, les écarts aux formes standard, le fait que l'écriture comporte une ou plusieurs sessions, qu'ils ont été ou non corrigés par l'enseignant, etc.

Régulièrement améliorée, la plateforme vise à devenir une ressource de référence pour la connaissance de l'écriture et de son apprentissage, de l'écriture débutante à la littéracie avancée, mais aussi pour la formation des enseignants. Une des suite du projet E-CALM est la collaboration avec le Centre Alain Savary de l'Institut Français de l'Éducation, à travers le projet OSERAIT (*Observer la Scripturalité chez les Elèves : Reflexivité et Analyses Interprétatives sur le Terrain*). Le Centre Alain Savary met à disposition des formateurs d'enseignants des outils et des scénarios de formation utilisant des résultats de recherche. À travers le projet OSERAIT sont testées des utilisations possibles des écrits scolaires dans les formations à la didactique de l'écriture permettant de produire des outils innovants pour la formation des enseignants, dans une double perspective actionnelle et réflexive. Six ans après son lancement, le projet E-CALM arrive donc à un palier, celui de la dissémination des résultats de la recherche et de leur opérationnalisation sur les terrains scolaire et universitaire.

Bibliographie

- Bellemin-Noël, J. (1979). Lecture psychanalytique d'un brouillon de poème : Été de Paul Valéry. In *Essais de Critique Génétique*, Flammarion.
- Boë L.-J. (1997). Sciences phonétiques et relations forme/substance : 1. un siècle de ruptures, négociations et réorganisations. *Histoire Épistémologie Langage*, tome 19, fascicule 1. 5-41. doi : <https://doi.org/10.3406/hel.1997.2571>.
- Brossard, M. (2013). Enseignement-apprentissage scolaire et développement : analyse d'un exemple In : *Vygotski et l'école : Apports et limites d'un modèle théorique pour penser l'éducation et la formation*. Pessac : Presses Universitaires de Bordeaux.
- Burnard, L. (2007). "New tricks from an old dog: An overview of tei p5". In Lou Burnard, et al., editors, *Digital Historical Corpora- Architecture, Annotation, and Retrieval*, number 06491 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany.
- Charolles, M. (1978). Introduction aux problèmes de la cohérence des textes. In *Langue française*, no 38, p. 7-42.
- Crasson, A., Fekete, J.-D. (2004). Structuration des manuscrits : Du corpus à la région. Semaine du Document Numérique (SDN 2004). Conférence Internationale Francophone sur l'Écrit et le Document (CIFED 04), Jun 2004, La Rochelle, France. 162-168. sic_00001210.
- Doquet, C. (2011). *L'Écriture débutante. Pratiques scripturales à l'école élémentaire*. Rennes, PUR.
- Doquet, C., Lefebvre, J., Mahrer, R., Testenoire, P.-Y. (2022). L'oral et l'écrit : si proches, si loin. *Linguistique de l'écrit 3*, Oral/écrit: Quelles places dans les modèles linguistiques?, 9-32. <https://doi.org/10.19079/lde.2022.s3.1>.

- Ivins, W. M. (1953). *Prints and Visual Communication*. A Da Capo Press reprint edition. Harvard University Press. <https://books.google.it/books?id=IQsnAAAAMAAJ>.
- Lebrave, J.-L. (1999). Les manuscrits entre la substance de l'expression et la substance du contenu. In C. Cortès et A. Rousseau (eds), *Catégories et connexions. En hommage à Jean Fourquet pour son centième anniversaire*, Villeneuve d'Ascq, Presses Universitaires du Septentrion.
- Lebrave, J.-L. (2020). *Théorie et linguistique de l'écriture. Des manuscrits aux processus scripturaux (1983-2018)*, B. Vauthier et R. Mahrer (éd.). Paris : Classiques Garnier. ISBN : 978-2-406-09341-1.
- Lumbroso, O., Mitterand, H. (2002). Les Manuscrits et les dessins de Zola : Notes préparatoires et dessins des Rougon-Macquart. Edition établie et commentée par Olivier Lumbroso et Henri Mitterand. Editions Textuel.
- Luria, A.R. (2002). Contribution à l'étude de la genèse de l'écrit chez l'enfant, traduit dans Apprendre à l'école : perspectives piagétienne et vygotkiennes, M. Brassard et J. Fijalkow (dir.), Bordeaux, PUB, 2002.
- Mahrer, R. (2014). Écrire et parler. Quelques préalables théoriques », *Genesis* [En ligne], 39. URL : <http://genesis.revues.org/1371> ; DOI : 10.4000/genesis.1371.
- Mellet, S. (2002). Corpus et Recherches Linguistiques: Introduction. *Corpus*, no. 1, <https://doi.org/10.4000/corpus.7>.
- Rastier, F. (2004). Enjeux épistémologiques de la linguistique de corpus. *Texte !* [en ligne]. Rubrique Dits et inédits. Disponible sur : <http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html>.
- Vygotskij, Lev Semenovič, et al. (2014). Histoire du développement des fonctions psychiques supérieures, La Dispute.