CORPUS

Corpus

16 | 2017
Spécificités et contraintes des grands corpus de textes scolaires : problèmes de transcription, d'annotation et de traitement

Un « corpus de littéracie avancée : résultat et point de départ

Marie-Paule Jacques et Fanny Rinck

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/2806
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Édition imprimée
Date de publication : 1 janvier 2017
ISBN : 16638-9808
ISSN : 1638-9808

Référence électronique
Marie-Paule Jacques et Fanny Rinck, « Un « corpus de littéracie avancée : résultat et point de départ », *Corpus* [En ligne], 16 | 2017, mis en ligne le 07 janvier 2018, consulté le 08 septembre 2020. URL : http://journals.openedition.org/corpus/2806

Un « corpus de littéracie avancée : résultat et point de départ

Marie-Paule Jacques et Fanny Rinck

1. Introduction

1 La constitution d'un corpus de littéracie avancée va de pair avec la constitution de la littéracie avancée comme champ. C'est ce que nous proposons de montrer dans cet article : à travers la structure du corpus, ses méta-données et les annotations possibles se jouent des principes fondateurs et évolutifs, qui sont étroitement liés au rôle des recherches linguistiques et des corpus en didactique de l'écrit, en l'occurrence dans l'enseignement supérieur. Nous traiterons donc de la dimension technique du corpus comme de ses utilisations présentes et potentielles, avec en ligne de fond des préoccupations d'ordre épistémologique sur les corpus en linguistique et en didactique.

2 Le corpus de littéracie avancée est pensé comme un résultat et un point de départ. Un corpus n'est pas une donnée brute mais suppose une construction. L'engouement pour les corpus peut conduire à des apories : leur construction est un préalable à leur exploitation, mais on court alors le risque qu'ils restent lettre morte. Que les chercheurs fassent des corpus est une chose, mais qu'en font-ils ? L'idée d'un produit abouti est à nuancer : loin d'être intangible, il est, comme l'écrit B. Habert (2005) à propos des seules annotations, « un résultat temporaire qu'on doit pouvoir corriger, faire évoluer ». En ce sens, nous le voulons aussi un point de départs multiples vers des recherches variées, des extensions, des enrichissements, de nouvelles perspectives.

3 Nous commencerons par préciser les objectifs linguistique et didactique qui motivent la constitution et la mise à disposition d'un tel corpus. Nous expliciterons ensuite les choix liés aux méta-données, pensés dans la perspective d'un outillage des linguistes et des didacticiens dans le champ de la littéracie avancée. Nous développerons enfin les potentialités offertes par ce corpus en tant qu'instrument de recherche et ressource pour la formation à l'écrit.

2. Quels objectifs pour le corpus ?

4 Le corpus de littéracie avancée s'inscrit dans la continuité d'une pratique qui existe de longue date, celle qui consiste pour les enseignants et formateurs de langue à collecter et utiliser des productions d'apprenants. Son enjeu, à la fois linguistique et didactique, est à comprendre aussi en lien avec l'intérêt des possibilités offertes aujourd'hui par le numérique.

5 Les objectifs du corpus couvrent les trois dimensions que citent A. Boulton et H. Tyne (2014 : 7), en référence aux fondateurs des « learners corpora », du « teaching/learning with corpora » et des « corpus linguistics in language teaching », tels S. Granger *et al.* (2002), J. Sinclair (2004) et C. Trible (1997) :

6 1) Les apprenants sources de données : corpus de productions d'apprenants au long de leur parcours.

7 2) Les apprenants bénéficiaires (utilisation dite indirecte) : corpus servant de ressources pour des supports pédagogiques : dictionnaires, manuels et, comme il en sera question ici, élaboration d'exercices.

8 3) Les apprenants utilisateurs des corpus (utilisation dite directe) : concept de « l'apprenant-chercheur » (voir aussi Boulton & Tyne, 2014 : 77), ou s'agissant d'apprentissages langagiers, de « détective linguistique » (Johns, 1997), et, à l'échelle des textes et discours, de « language learning as discourse analysis » (Chambers, 2007).

9 Dans ce qui suit, nous détaillons d'abord les objectifs d'ordre descriptif, qui correspondent au premier item : les apprenants comme source de données. Nous en viendrons ensuite au corpus en tant que ressource pour la formation (2.2).

2.1 Pour l'analyse empirique de la littéracie avancée

10 Un premier objectif d'ordre descriptif est de cerner les compétences rédactionnelles et d'identifier les besoins de formation à un niveau avancé. La désignation de « littéracie avancée » se base sur l'idée d'apprentissages de l'écrit en continu et se justifie par le fait qu'il s'agit de pointer un niveau de compétences plutôt qu'un contexte ou une visée communicative spécifique. L'enjeu est d'éclairer 1) les acquis et les difficultés à l'écrit chez des publics dits avancés d'étudiants ou de cadres, identifiés dans l'enseignement supérieur mais aussi dans le monde professionnel et 2) les études sur l'expertise rédactionnelle et l'écriture de haut-niveau, autrement dit sur la possibilité d'aller toujours plus loin dans le développement de compétences, face à de nouvelles pratiques.

11 L'orthographe retient prioritairement l'attention, mais elle n'est pas la seule concernée : la construction des phrases, les choix lexicaux, la cohérence textuelle le sont aussi, comme, dans une dimension pragmatique, l'adaptation au genre et au destinataire. Il faut en ce sens rappeler que l'enjeu n'est pas seulement de bien écrire, mais plutôt de penser et d'agir à travers l'écrit.

12 Le constat fait par M.-L. Elalouf et C. Boré (2007) concernant la rareté des corpus d'écrits scolaires en France et en français reste vrai. Or ce sont bel et bien les outils de la linguistique de corpus qui permettront d'aller au-delà de constats que tout un chacun peut faire. La caractérisation linguistique des pratiques effectives en littéracie

avancée n'en est qu'à ses débuts et mérite d'être systématisée pour prendre en compte les différentes dimensions de la maîtrise de l'écriture et dans différents genres.

13 Il faut aussi envisager de manière dynamique les acquisitions tardives et se demander en quoi les pratiques évoluent, entre la première année de licence et la 2e année de master. Par rapport aux niveaux antérieurs de scolarité, on peut s'appuyer sur les travaux des didacticiens de l'écrit qui adoptent cette même perspective d'analyse de corpus et de caractérisation linguistique des productions, comme C. Boré (2007), D. Cogis (2013), J. David (ex. David & Marin, 2013), C. Doquet (2011), M.-L. Elalouf (ex. Elalouf & Boré, 2007, Elalouf *et al.*, 2012), C. Garcia-Debanc (ex. Garcia-Debanc & Bonnemaison, 2014), M.-J. Béguelin, pionnière en la matière et spécifiquement à propos des lycéens et étudiants ([Reichler]-Béguelin *et al.*, 1988). L'enjeu est de contribuer à l'analyse des apprentissages de l'écriture en France et en français de la maternelle à l'université : quels sont les apprentissages tardifs ? Peut-on parvenir à identifier des étapes entre des productions peu normées et des productions satisfaisant aux attentes ? Peut-on ainsi aboutir à des référentiels de compétences ? Le processus n'est sans doute pas uniquement linéaire : des difficultés nouvelles apparaissent avec les genres nouveaux auxquels le scripteur adulte est confronté, et plus largement des difficultés spécifiques existent selon les genres considérés : c'est réaffirmer qu'il existe une grammaire des genres.

14 Par ailleurs, les textes des étudiants ne sont pas à interpréter qu'en termes de défaillances ; par exemple la structure de leurs phrases rejoint des analyses faites dans la presse sur les figures d'ajout après le point, qui signalent des évolutions dans les usages de l'écrit (Boch *et al.*, 2015).

15 Le corpus de littéracie avancée offre aussi un point de comparaison utile par rapport aux études en langue étrangère (FLE) ou langue seconde (FLS), auxquelles manque une appréciation de ce que des scripteurs natifs d'un âge et d'un niveau d'étude similaires sont en mesure de produire. Le domaine du FOU (Français sur Objectifs Universitaires) (Mangiante & Parpette, 2011) qui émane de celui du FLE, peut représenter ce lieu fédérateur où sont mises en perspective les difficultés rédactionnelles en français des étudiants, en fonction de leur niveau en français et de leur biographie langagière (voir notamment Hatier & Yan 2015). Le corpus de littéracie avancée poursuit donc une visée similaire à System[1]) qui se centre sur les productions enfantines à l'oral dans une perspective plurilingue, fournit des normes de transcription et des outils pour la transcription et l'analyse et représente une riche base de données pour des recherches à l'échelle internationale.

2.2 Comme ressource pour les formateurs et les apprenants

16 Le second objectif du corpus de littéracie avancée relève d'une exploitation didactique du corpus. Il constitue une ressource pour la formation, à la fois pour l'élaboration de supports et activités et pour une utilisation directe par les apprenants, selon le principe présenté plus haut (Boulton & Tyne, 2014) : courts extraits ou textes peuvent faire l'objet d'observations ou de manipulations et favoriser le raisonnement, dans une perspective aussi bien de remédiation que de formation de spécialistes (rédacteurs, correcteurs, formateurs).

17 Nous avons expérimenté cette utilisation à travers un projet de type « PedagoTice » (Jacques & Rinck, 2015), mettant en application le principe d'appui sur les textes

d'apprenants développé par P. Cappeau & M.-N. Roubaud (2005), à travers la construction d'exercices ciblés.

18 Les textes du corpus mettent en évidence réussites et maladresses de leurs auteurs. Les réussites ont ceci de remarquable qu'elles permettent de restituer aux apprenants des formulations, des passages qui ne présentent pas le caractère souvent ressenti comme inaccessible de l'écriture experte, mais des solutions à leur portée, élaborées par leurs pairs, qui peuvent alors jouer pleinement le rôle de *modèle*. Le contraste avec des passages similaires non réussis (par exemple des extraits d'introduction, des passages censément explicatifs ou argumentatifs), le recours à des jugements d'acceptabilité alimentent une réflexion sur ce qui fait que « ça fonctionne » ou « ne fonctionne pas ».

19 Il ne s'agit pas de demander à l'apprenant de simplement « trouver la réponse juste », mais d'aiguiser son regard et de l'entraîner ainsi à une réflexivité essentielle dans le développement de ses compétences scripturales.

20 Par exemple, on fait comparer des amorces d'introduction, comme :

21 - « La poésie est un genre littéraire important. Le corpus proposé réunit trois auteurs du 20e siècle qui pratiquent la poésie et se questionnent sur leur pratique. »

22 - « La poésie n'est pas que l'art de réciter ou d'écrire en rimes et en vers. C'est ce que montre le corpus proposé : il réunit trois auteurs du 20e siècle qui pratiquent la poésie et se questionnent sur leur pratique. »

23 Dans les deux amorces, le corpus de textes est présenté, l'enjeu pointé (la poésie en acte et en question). La première s'ouvre par une affirmation générale sur l'importance du sujet, qui vaudrait pour tout genre, la seconde part de l'idée spontanée que l'on peut se faire de la poésie.

24 Ce même principe de comparaison est utilisé dans des exercices sur la syntaxe ou le lexique. Par exemple, on demande aux étudiants quel énoncé leur semble plus adapté :

25 - « Prévert met en scène son animal favori, l'oiseau, pour expliquer comment faire un poème, il faut d'abord "peindre une cage" ».

26 - « Prévert met en scène son animal favori, l'oiseau, pour expliquer comment faire un poème : il faut d'abord "peindre une cage" ».

27 Les exerciseurs utilisés (HotPotatoes, Opale Sup de Scenari, Chamilo LMS) offrent diverses possibilités : cliquer sur l'énoncé le plus adapté, ou sur l'élément défaillant, ajouter ou non une lettre pour l'accord dans un espace laissé vide, cliquer sur un onglet « indice » ou sur la « solution ».

28 En présentant des phrases ou extraits jugés défectueux assortis de propositions de réécriture, on met donc en évidence les zones de révision qui peuvent aller de la correction orthographique très locale au remaniement complet d'une phrase « bancale ». Comme indiqué précédemment, la particularité de la démarche réside dans le lien étroit entre exercices et genre de texte. Il ne s'agit pas d'aider les étudiants à acquérir des connaissances générales sur la langue mais bien de les aider à progresser dans la maitrise de formes langagières requises par des exigences textuelles propres au monde académique et nouvelles pour eux.

29 À titre de comparaison, certains des exercices que propose le *Centre Collégial de Développement de Matériel Didactique (CCDMD) – Amélioration Du Français*[2] portent sur une connaissance fine de contraintes linguistiques du français, mais non ciblée sur les écrits

académiques : par exemple repérage de problèmes d'accord dans le cas de structures Nom Nom telles que *voyages /éclair/, pâtisseries /maison/* ou *appartements /témoin/*.

30 Le corpus de littéracie avancée rend *a contrario* possible la focalisation non sur un savoir encyclopédique dont on ne sait pas s'il est rentable pour les étudiants mais sur des difficultés récurrentes. Un bon exemple est celui de la formulation d'une problématique, dans une introduction de dissertation ou dans un mémoire. L'exigence se situe au niveau des choix terminologiques mais l'on voit aussi quantité de scripteurs trébucher sur la mise en mots de l'interrogation indirecte qui requiert maitrise tout à la fois de l'ordre des mots (pas d'inversion du sujet), du choix du couple verbe-subordonnant, de la ponctuation.

31 Le corpus permet de collecter de façon systématique les énoncés défaillants, soit par une lecture extensive, soit en sélectionnant uniquement certains textes, en exploitant alors les métadonnées qui leur sont associées.

3. Méta-données : pour un outillage des linguistes et des didacticiens dans le champ de la littéracie avancée

32 Les objectifs précédemment évoqués envisagent diverses pistes d'exploitation du corpus ; nous développerons dans la section 4 d'autres potentialités qui ne prennent sens qu'à concevoir ce corpus comme un construit évolutif, stabilisé ponctuellement en un certain état, mais susceptible d'intégrer de nouvelles données pour une évolution vers un état enrichi.

33 En janvier 2016, le corpus contient 338 textes (+ d'1 million de mots) produits par des étudiants en France et en français, à des niveaux variés, de la licence première année au master 2e année. Diverses disciplines sont représentées ainsi que divers genres allant des écrits universitaires (fiches de lecture, mémoires) à des écrits professionnels, dans le cadre de formations de rédacteurs et formateurs à l'écrit (lettres de motivation, comptes-rendus)[3].

34 Cette diversité qui a présidé à la constitution du corpus ne répond pas à une ambition de représentativité de la diversité de l'écrit académique (mettre un peu de tout) mais à la volonté de bâtir une base au sein de laquelle puisse s'opérer une sélection cohérente avec des objectifs de recherche non nécessairement programmés dès l'origine. Cette constitution s'inscrit dans la lignée de bases existantes telles que Frantext ou Scientext[4], qui autorisent la confection à volonté de « corpus d'étude » (Rastier, 2005).

35 Il est alors essentiel d'anticiper d'une part les critères de sélection des textes, d'autre part les évolutions possibles du corpus pour recevoir de nouveaux enrichissements. Le premier point se traduit par la documentation de métadonnées, le second par un choix de formatage xml.

36 La notion de méta-données, indispensables à la production d'un corpus informatisé, recouvre en réalité deux sous-ensembles d'informations à associer aux textes du corpus :

37 - des précisions sur l'élaboration du corpus – qui a rassemblé les textes, étaient-ils originellement informatisés ou ont-ils été saisis ou numérisés, sont-ils publics, publiés, traduits, etc. ?

38 - la documentation et caractérisation de chacun des textes et des sous-corpus.

39 Si les items du premier sous-ensemble sont désormais bien identifiés pour peu que l'on se réfère à des recommandations telles que celle élaborée par la Text Encoding Initiative (TEI)[5], le second sous-ensemble, en tant qu'il renseigne nature et contenu, est le produit de choix et décisions non fixés à l'avance, répondant à la nécessité d'offrir un aperçu synthétique du texte, de ce qui le caractérise et de ce qu'il partage dans son sous-ensemble. Il constitue la description de son profil (*profileDesc* dans la terminologie de la TEI). Pour illustration, la Figure 1 montre un extrait de l'encodage des méta-données au format XML-TEI.

Figure 1. Extrait de l'en-tête contenant les méta-données

40 Les items choisis pour cette caractérisation résultent de la prise en compte des besoins des chercheurs en linguistique et didactique tels qu'ils se déduisent des recherches. Ces items représentent autant de variables usuellement convoquées soit comme variable indépendante, soit comme variable dépendante de la recherche.

41 Chaque texte est ainsi accompagné d'une indication du niveau d'études de son auteur, codifié de façon simple en *L1 L2 L3*, *M1 M2*. De même, la « discipline » est explicitée, non de façon standardisée, mais en employant les expressions mêmes choisies par l'université pour intituler le diplôme. Par exemple, tel texte viendra du M1 « Formateurs d'enseignants 1er et 2nd degré », tel autre de L3 « Didactique du français ». Ces éléments sont complétés par le nom de l'université et l'année universitaire de production.

42 La langue des textes est aussi précisée, quoiqu'elle soit ici uniforme : français pour tous. Un item est prévu pour la langue maternelle de l'auteur, mais jusqu'ici cette information n'a pas été consignée lors du recueil du texte et l'item reste donc non renseigné.

43 Plusieurs items concernent la nature du texte et ses conditions de réalisation. Le genre du texte est ici une donnée incontournable, mais qui pose évidemment de réels problèmes de définition. Il n'existe pas d'inventaire *a priori* et consensuel des genres académiques[6], bien que les *curricula* comportent des productions écrites peu ou prou similaires : quelle que soit la disci-pline, un étudiant peut s'attendre à réaliser une *fiche de lecture*, un *rapport de stage*, un *dossier*, un *mémoire*... Même dans le cas d'une dénomination commune, les exigences quant à la longueur, au contenu et à la teneur de l'écrit attendu sont variées. La solution la plus transparente consiste là encore à inscrire comme genre la dénomination produite par l'institution elle-même. On trouvera donc dans le corpus des *compte-rendus de lecture*, des *synthèses théoriques* à côté de *mémoires* et *compte-rendus professionnels.* Il s'agit bien du genre prescrit et non d'une caractérisation *a posteriori* du texte en fonction de propriétés qui le feraient appartenir à tel ou tel genre (Malrieu & Rastier, 2001). Des études peuvent alors être menées sur les propriétés formelles d'un genre tel que les étudiants se l'approprient et sur leurs difficultés par rapport aux propriétés attendues.

44 En complément de cette indication de genre, est reportée la consigne d'écriture du texte. Il s'agit là encore d'une information que les chercheurs peuvent mobiliser soit comme préalable à une étude par exemple contrastive d'un même genre mais de consignes variées, soit comme variable explicative des observations qui peuvent être menées sur le-s texte-s.

45 De même, avec chaque texte est précisé le nombre d'auteurs. Peu de textes sont écrits « à plusieurs mains », dans la mesure où les cursus universitaires s'appuient massivement sur les productions individuelles pour la diplomation, huit seulement sont le fait de deux auteurs, mais ils peuvent amorcer des études contrastives pour mesurer les effets de ce facteur.

46 Deux derniers items liés aux conditions de production concernent le nombre de versions et le numéro de version du texte. Certains textes sont en effet une réécriture d'une version précédente. Les utilisateurs du corpus peuvent ainsi s'ils le souhaitent interroger ces textes via un alignement de ces versions (voir plus loin, note 8).

47 Ajoutons à ce qui précède que la taille en nombre de mots est indiquée pour chaque texte, et l'on a un tableau complet des éléments qui sont à la fois supposés pertinents et informatifs et réalistement récoltables.

48 Ils peuvent toutefois être considérés pour certaines études comme trop pauvres en informations contextuelles, c'est-à-dire en « données extratextuelles » explicitant « le cadre institutionnel dans lequel [les textes] sont produits et les contraintes discursives et institutionnelles qui leur sont imposées » (Cislaru & Sitri, 2009 : 85). On ne sait pas de façon systématique quels travaux préparatoires les ont précédés le cas échéant, ni quel est le bagage des auteurs (quel baccalauréat, quelle licence ?), ni quels sont leur âge et vécu antérieur, ni le moment de l'année où se situe le travail, autant d'éléments qui pourraient entrer en compte pour l'interprétation des faits observés lors d'une recherche. Mais c'est le corollaire inévitable du passage d'une échelle propice aux analyses qualitatives menées sur des données richement documentées à une échelle propice à l'extraction de régularités et à l'émergence de contrastes entre ensembles. Si certains sous-ensembles ne se saisissent qu'avec une part de contextualisation (une fiche de lecture exige une mise en relation avec le texte primaire lu, par exemple, d'où la présence des consignes qui contiennent cette référence), le corpus pris globalement n'exige pas cette contextualisation systématique.

49 Car le corpus est pensé non pour étayer des analyses quant à l'impact de certaines situations didactiques sur la qualité des productions, mais pour permettre une approche à gros grain des compétences rédactionnelles d'une population saisie dans son ensemble. Il est voué à une intégration dans des logiciels d'exploration de texte[7] qui prévoient une sélection libre d'un corpus de travail, pour laquelle les métadonnées exposées (ex. genre, niveau, discipline) prennent leur sens.

50 L'adoption d'un format xml offre l'intérêt majeur de permettre cette intégration et de garantir une certaine interopérabilité, renforcée par le cadre d'un schéma TEI.

51 Ce standard fournit pour les métadonnées des champs génétiques tels que « profileDesc » et « textClass » que l'on décline ensuite à notre guise, en sous-champs appropriés à notre corpus, comme dit plus haut. Balises qui permettent de repérer dans les textes mêmes soit des éléments de structure, tels que paragraphes ou sections, soit des éléments discursifs tels qu'une citation. La version actuelle du corpus (janvier 2016) ne fait usage que des balises structurelles et ne comporte aucune balise qui apporterait des éléments d'analyse du contenu (par ex. sur les erreurs et maladresses des étudiants).

4. Potentialités du corpus pour la recherche et la formation à l'écrit

52 On aura compris que la mise à disposition du corpus de littéracie avancée repose sur un principe de mutualisation ; elle permet de faire en sorte qu'au-delà des recherches que peut mener une équipe sur un corpus constitué isolément, tout le champ de la littéracie avancée puisse y avoir recours. Le corpus a donc été pensé en regard des potentialités qu'il offre et qu'il pourra offrir face à des objectifs variés et évolutifs. Par définition, ces potentialités ne sont pas pré-établies, mais certaines cependant sont d'ores et déjà bien identifiées et concernent d'une part des extensions du corpus pour de nouvelles recherches sur les pratiques scripturales et d'autre part des exploitations didactiques.

4.1 Des données sur les processus d'écriture

53 Une des premières extensions envisagée est d'intégrer au corpus de textes des données sur le processus d'écriture. En l'état actuel, on dispose pour quelques textes de deux voire trois versions qui sont autant d'étapes d'écriture. Comme le détaille C. Doquet à propos de l'écriture débutante (2011), les données génétiques sont précieuses car les ratures et opérations de réécriture sont la trace d'une interrogation méta-discursive spontanée (Fabre, 1990, 2002) elle-même au fondement de la normalisation des écrits. À partir de plusieurs versions d'un texte, on peut viser une caractérisation linguistique en prise avec les opérations de suppression, ajout, remplacement, déplacement, pour voir quelles unités linguistiques concernent ces opérations ou si telle unité préalablement analysée comme une difficulté, par exemple la ponctuation ou les accords morphographiques, fait l'objet de réécritures[8].

54 Le champ des études génétiques va cependant au-delà de l'analyse de versions successives d'un texte et se base aujourd'hui sur des données dites d'écriture enregistrée (Leblay & Caporosssi, 2014) : elles représentent l'écriture en acte et permettent ainsi de croiser sa dimension spatiale (topographie) avec sa dimension

temporelle (chronologie). L'idée est d'aboutir à un alignement entre un film d'écriture, un codage des opérations (ajout simple, suppression, modification) et l'état du texte à un moment donné. On peut alors, par exemple, analyser les difficultés rédactionnelles sous l'angle double des textes produits et des hésitations du scripteur rendues tangibles par le temps qu'il passe et/ou les modifications qu'il porte à son texte.

4.2 Des textes annotés par leurs relecteurs-correcteurs

55 Une autre extension à apporter au corpus de littéracie avancée consiste à intégrer des textes annotés par des relecteurs-correcteurs[9], en particulier les enseignants à l'origine de la tâche de production écrite.

56 Les annotations peuvent être de deux types : des modifications dans le texte et des commentaires sur le texte[10]. Elles peuvent consister à identifier des erreurs ou maladresses ou porter sur des passages plus larges (par exemple des commentaires comme « à reformuler », « trop oral », « bien »).

57 De telles données permettent d'envisager la dimension collaborative de l'écriture : dans le champ de la littéracie avancée, il est crucial de considérer, au-delà de l'interaction didactique, les relectures faites par les pairs et l'écriture à plusieurs mains (très présente dans le monde professionnel). Par ailleurs, ces données pourraient permettre de mieux cerner les normes aux fondements des pratiques scripturales et la diversité dans les attentes des lecteurs : à la manière de l'étude de J.-L. Pilorgé (2008) sur les postures du lecteur-correcteur des écrits scolaires, qu'en est-il pour la littéracie avancée ? A-t-on affaire à des relecteurs intrusifs, qui modifient le texte en plus de le commenter ? Quelles sont les zones de consensus entre relecteurs et les zones d'appréciation plus subjective ? Par exemple, quelles interventions sont faites concernant la ponctuation ? La cohérence argumentative ? Un large champ d'étude peut alors s'ouvrir, qui rejoint d'une part les travaux sur la manière dont les éditeurs interviennent sur les textes et pose d'autre part la question de l'acceptabilité et de la frontière entre ce qui relève de la correction linguistique et ce qui relève plutôt de considérations stylistiques.

4.3 Des dispositifs pour la formation à l'écrit

58 Les potentialités du corpus évoquées jusqu'ici répondent à des objectifs de recherche linguistique. Concernant ses exploitations pour la formation à l'écrit, les premières réalisations ont consisté à repérer des passages intéressants, soit parce qu'ils présentent des erreurs ou maladresses, soit au contraire parce qu'ils sont réussis, et à les utiliser dans des exercices et activités.

59 Il est important d'abord de penser à la manière dont on peut faciliter l'accès du formateur à des passages intéressants du corpus. Un outil de requête adapté s'impose, pour pouvoir sélectionner dans un genre spécifique (ou un ensemble de genres), pour un niveau spécifique (ou à des niveaux différents), des extraits de textes correspondant aux objectifs didactiques (ex. faire travailler les étudiants sur des introductions de mémoire, sur la citation, sur les connecteurs argumentatifs, etc.). Dans son état actuel, le corpus peut être associé à un outil de requête permettant de sélectionner un sous-ensemble du corpus en fonction des méta-données présentées dans la partie 3 (genre, niveau, etc.). Pour aller plus loin, il faut envisager des annotations concernant la

structure des textes (introduction, citation) ainsi que les types de difficultés rédactionnelles (orthographe, lexique, connecteurs, etc.). Le problème est de savoir quoi et jusqu'où annoter, et comment s'assurer que ce qui relève de la construction du corpus (en l'occurrence, ses enrichissements) soit en phase avec les besoins du formateur. L'annotation ouverte et collaborative des documents représente une perspective prometteuse : diverses équipes pourraient peu à peu contribuer à enrichir le corpus au fur et à mesure de recherches sur telle ou telle difficulté rédactionnelle.

60 Cette utilisation du corpus par les formateurs, dite indirecte par A. Boulton et H. Tyne (2014) (cf. début de la section 2) rejoint celle qu'ils qualifient de directe : l'observation et l'analyse d'extraits du corpus par les apprenants, à l'aide de questions de guidage conçues par l'enseignant en référence aux objectifs didactiques. Dans l'enseignement des langues étrangères, les concordanciers ont typiquement été utilisés dans ce sens.

61 L'observation et l'analyse peuvent concerner non pas seulement tel ou tel passage (ex. concordances) mais la mise en relation d'unités textuelles, par exemple pour étudier une anaphore en lien avec le passage qu'elle reprend, ou pour étudier le marquage d'une structure énumérative ou argumentative dans un large extrait de texte. C'est ce principe qui est au fondement de la navigation didactique dans les textes (Lundquist, 2008, 2013) et du logiciel Navilire (Couto, Lundquist, Minel, 2005), conçu pour l'aide à la production de textes dans le cadre de l'apprentissage des langues étrangères. Il vise à former les étudiants à la question de la cohérence textuelle en leur demandant de naviguer dans des textes pour identifier les unités assurant la cohérence et leurs relations. Les annotations préalablement portées sur les textes par les formateurs permettent aux étudiants d'obtenir un feed-back (questions de guidage, corrections) au cours de l'exercice. On peut envisager de développer ce type de logiciel à partir des logiciels d'annotation à l'échelle textuelle comme Glozz (Widlöcher & Mathet, 2012), Analec (Landragin *et al.*, 2012), basé sur le même principe des annotations dites textuelles ou discursives[11].

62 En somme, les potentialités offertes par le corpus de littéracie avancée se situent pour certaines à court terme, d'autres à plus long terme et s'adressent aux utilisateurs multiples que sont les linguistes et chercheurs du domaine de la littéracie avancée, les formateurs et les apprenants. Le corpus (ne) représente donc (qu') un point de départ auquel sont à associer 1) des outils de requête, à la manière des bases Frantext ou Scientext (voir note 4), pour la sélection par le chercheur ou par le formateur d'extraits de texte correspondant à telle opération de réécriture et/ou à tel type de difficulté rédactionnelle, et 2) des outils d'annotation adaptés (ou détournés) selon les utilisations : annotations de haut-niveau, comme sur les anaphores, comparaison d'annotations faites par plusieurs lecteurs sur le style, exercices consistant pour les apprenants à annoter et à recevoir en retour un guidage ou une correction.

5. Conclusion

63 Le corpus de littéracie avancée synthétise d'une certaine manière les avancées du champ auquel il emprunte son nom. Il prolonge la logique des études menées sur les littéracies universitaires, qui font émerger l'importance des genres pour les compétences rédactionnelles (Delcambre & Lahanier, 2010)[12]. Il est conçu de façon évolutive et modulable pour servir non seulement les objectifs visés originellement

mais une variété de recherches allant du local[13] au contrastif – en mobilisant éventuellement d'autres corpus – dans plusieurs directions :

64 - développement (étude d'une compétence donnée du primaire à l'université) ;

65 - interlangue (étude d'une compétence donnée pour un natif *vs* apprenant non natif) ;

66 - intergenre (étude d'un même phénomène linguistique dans plusieurs genres) ;

67 - ...

68 Son format XML, dédié au partage, le rend compatible avec la panoplie d'outils logiciels adaptés à la diversité des recherches et des objectifs envisagés et, nous l'espérons, non encore imaginés.

BIBLIOGRAPHIE

Boch F., Cavalla C., Pétillon S. & Rinck F. (2015). « Travailler le texte : ponctuation, anaphores, collocations », *in* F. Boch & C. Frier (éd.) *Écrire dans l'enseignement supérieur : des apports de la recherche aux outils pédagogiques*. Grenoble : ELLUG, 53-109.

Boré C. (éd.) (2007). *Construire et exploiter des corpus de genres scolaires, Diptyque* 10. Namur : Presses Universitaires de Namur.

Boulton A. & Tyne H. (2014). *Des documents authentiques aux corpus. Démarches pour l'apprentissage des langues.* Paris : Didier.

Cappeau P. & Roubaud M.-N. (2005). *Enseigner les outils de la langue avec les productions d'élèves : cycles 2 et 3*. Paris : Bordas.

Chambers A. (2007). « Language learning as discourse analysis : Implications for the LSP learning environment », *ASp* 51-52 : 35-51. En ligne, http://asp.revues.org/483.

Cislaru G. & Sitri F. (2009). « Texte et discours : Corpus, co-texte et analyse automatique du point de vue de l'analyse de discours », *Corpus* 8 : 85-104.

Cogis D. (2013). « La révision orthographique au CM2 : l'accord sujet-verbe dans le corpus Grenouille », *in* C. Gunnarsson-Largy C. & E. Auriac-Slusarczyk (éd.) *Écriture et réécritures chez les élèves. Un seul corpus, divers genres discursifs et méthodologies d'analyse*. Louvain-la-Neuve : Academia, 85-112.

Couto J., Lundquist L. & Minel J.L. (2005). « Naviguer dans les textes pour apprendre », *Actes de TALN 2005*, Dourdan. En ligne, http://halshs.archives-ouvertes.fr/docs/00/09/80/24/PDF/TALN_COUTO_LUNDQUIST_MINEL.PDF.

David J. & Marin B. (éd.) (2013). Écrits d'élèves, contraintes de la langue, *Le Français Aujourd'hui* 181. Paris : Larousse.

Delcambre I., Lahanier-Reuter D. (2010). « Les littéracies universitaires : Influence des disciplines et du niveau d'étude dans les pratiques de l'écrit », *Diptyque* 18 : 11-42.

Doquet C. (2011). *L'écriture débutante. Pratiques scripturales à l'école élémentaire.* Rennes : PUR.

Elalouf M.-L. & Boré C. (2007). « Construction et exploitation de corpus d'écrits scolaires », *Revue française de linguistique appliquée* 12 : 53-70.

Elalouf M.-L., Beaumanoir-Secq M., Bornaz S., Fort P.-L. (2012). « Enjeux de la constitution de corpus dans les écrits professionnels et de recherche du master "éducation et formation" : le cas de la didactique du français », *in* M.-L. Elalouf, A. Robert, A. Belhadjin & M.-F. Bishop (éd.) *Les didactiques en question*. Bruxelles : De Boeck, 382-403.

Fabre C. (1990). *Les brouillons d'écoliers ou l'entrée dans l'écriture.* Grenoble : Ceditel.

Fabre C. (2002). *Réécrire à l'école et au collège.* Paris : ESF.

Falaise A., Tutin A. & Kraif O. (2011). « Définition et conception d'une interface pour l'exploitation de corpus arborés pour non-informaticiens : la plateforme ScienQuest du projet Scientext », *TAL* 52-3 : 103-128.

Garcia-Debanc C. & Bonnemaison K. (2014). « La gestion de la cohésion textuelle par des élèves de 11-12 ans : réussites et difficultés », *Actes du Congrès Mondial de Linguistique Française (CMLF 2014)*, Berlin, Germany, 961-976.

Granger S., Hung J. & Petch-Tyson S. (éd.) (2002). *Computer Learner Corpora, Second Language Acquisition, and Foreign Language Teaching*. Amsterdam : John Benjamins.

Habert B. (2005). « Portrait de linguiste(s) à l'instrument », *Texto!* 10(4). En ligne, http://www.revue-texto.net/Corpus/Publications/Habert/Habert_Portrait.html.

Hatier S. & Yan R. (2015). « Comparaison de constructions verbales entre un corpus d'apprenants et un corpus d'articles de recherche », Communication orale, *8es Journées Internationales de Linguistique de Corpus* (*JLC 2015*), Orléans, 2-4 septembre 2015.

Ho-Dac L.-M. & Péry-Woodley M.-P. (2014). « Annotation des structures discursives : l'expérience ANNODIS », *Actes du Congrès Mondial de Linguistique Française (CMLF 2014)*, Berlin, Germany, 2647-2661.

Jacques M.-P. & Rinck F. (2015). « Une linguistique fondamentale et appliquée à base de corpus », Communication orale, *TRELA (Terrains de Recherche en Linguistique Appliquée)*, Université Paris Diderot, Paris, 8-10 juillet 2015.

Johns T. (1997). « Contexts : the background, development and trialling of a concordance-based CALL program », *in* Wichman *et al.* (éd.) *Teaching and language corpora.* Harlow : Longman, 100-115.

Landragin F., Poibeau T. & Victorri B. (2012). « ANALEC : a New Tool for the Dynamic Annotation of Textual Data. European Language Resources Association (ELRA) », *International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey, 357-362.

Lardilleux A., Fleury S. & Cislaru G. (2013). « Allongos : Longitudinal Alignment for the Genetic Study of Writers' Drafts », *Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science* 7817 : 537-548.

Leblay C. & Caporossi G. (2014). *Temps de l'écriture. Enregis-trements et représentations*. Louvain-La-Neuve, Academia.

Libersan L., Claing R. & Fouchambert D. (2010). *Stratégies d'écriture dans les cours de la formation spécifique. Rapport 2009-2010*. En ligne, http://www.ccdmd.qc.ca/-media/doc_theo_div_Rapport_Formation_specifique.pdf.

Lundquist L. (2008). *Navigating in Foreign Language Texts*. Frederiksberg : Sam- fundslitteratur.

Lundquist L. (2013). *Lire un texte académique en français*. Paris : Ophrys.

Malrieu D. & Rastier F. (2001). « Genres et variations morphosyntaxiques », *Traitement Automatique des Langues* 42(2) : 547-577.

Mangiante J.-M. & Parpette C. (2011). *Le français sur objectif universitaire.* Grenoble : PUG.

Pilorgé J.-L. (2008). *Un lieu de tension entre posture de lecteur et posture de correcteur : les traces des enseignants de français sur les copies des élèves.* Thèse de doctorat, Université Rennes 2.

Rastier F. (2005). « Enjeux épistémologiques de la linguistique de corpus », *in* G. Williams (éd.) *La linguistique de corpus*. Rennes : PUR, 31-45.

[Reichler-]Béguelin M.-J., Denervau M. & Jespersen J. (1988). *Écrire en français. Cohésion textuelle et apprentissage de l'expression écrite*. Neuchâtel-Paris : Delachaux et Niestlé.

Sinclair J. (éd.) (2004). *How to Use Corpora in Language Teaching*. Amsterdam : John Benjamins.

Tribble C. (1997). « Improvising Corpora for ELT : Quick-and- dirty Ways of Developing Corpora for Language Teaching », *in* B. Lewandowska-Tomaszczyk & J. Patrick (éd.) *Practical Applications in Language Corpora. The Proceedings of PALC '97*. Lodz : Lodz University Press. http://www.ctribble.co.uk/text/Palc.htm.

Widlöcher A. & Mathet Y. (2012). « The Glozz Platform : A Corpus Annotation and Mining Tool », *Proceedings of the 2012 ACM Symposium on Document Engineering*, 171-180.

ANNEXES

Annexes : descriptions du corpus

Niveau	Genre	Nbre de textes	Nbre de mots	Discipline-Filière	Établissement
L2	Dossier	10	28094	Sciences du Langage	Université Grenoble 3
L3	Parties théoriques de rapports de stage	15	70153	Didactique du français (Licence Sciences du Langage)	Université Grenoble 3
M1	Parties théoriques de mémoires	10	35306	Didactique du français	Université Grenoble 3
M1	Analyse (sujet type CRPE)	69	91991	Enseignement 1er degré (Français)	ESPE de Grenoble
M1	Analyse (sujet type CRPE)	97	117479	Enseignement 1er degré (Français)	ESPE de Grenoble
M1	Compte- rendu professionnel	22	4828	Sciences du Langage - Master Ecrifore (Écriture, Formation, Remédiation)	Université Paris Ouest Nanterre La Défense

M1	TER	25	274355	Enseignement Éducation Médiation 1er degré	IUFM de l'Académie de Paris
M1 et M2	Mémoires	41	395441	Didactique du français	Université Grenoble 3
M1 et M2	Lettre de motivation	20	6045	Sciences du Langage - Master Ecrifore (Écriture, Formation, Remédiation)	Université Paris Ouest Nanterre La Défense
M2	Synthèses théoriques	10	15678	Didactique du français	Université Grenoble 3
M2	Compte- rendu de lecture	10	35692	Éducation (Master Éducation-Médiation 1er degré)	Université de Cergy-Pontoise, IUFM
M2	Compte- rendu de lecture	10	6314	Enseignement Éducation Médiation 1er degré	IUFM de l'Académie de Paris
Total		339	1081376		

Niveau	Genre	Consignes
L2	Dossier	Dossier argumentatif sur un corpus choisi parmi ceux étudié en cours : expliquer en quoi il est rhétorique.
L3	Parties théoriques de rapports de stage	Parties théoriques de rapports de stage sur l'enseignement/ apprentissage du français
M1	Parties théoriques de mémoires	Parties théoriques de mémoires qui comprendront une partie d'analyse de terrain
M1	Analyse (sujet type CRPE)	Sujet de type concours CRPE (épreuve de français) : à partir du corpus proposé (3 textes), vous analyserez comment des poètes du 20ème siècle ont défini leur travail et leur art.
M1	Analyse (sujet type CRPE)	Sujet de type concours CRPE (épreuve de français) : Vous analyserez les quatre textes du corpus en montrant quelle réalité le théâtre donne à voir aux spectateurs.
M1	Compte-rendu professionnel	Dans le cadre d'un cours de pratique des écrits professionnels, produire un compte-rendu synthétique à partir d'un genre oral polyphonique de type table-ronde (thèmes : l'école, la prévention routière)

M1	TER	Rédaction d'un mémoire de trente pages sur l'écriture et ses apprentissages ou la diversité linguistique des élèves
M1 et M2	Mémoires	Mémoires de didactique du français comprenant une partie théorique et des observations de terrain.
M1 et M2	Lettre de motivation	Dans le cadre d'un cours de pratique des écrits professionnels, produire une lettre de motivation pour une candidature de stage (candidature spontanée ou réponse à une offre choisie par l'étudiant)
M2	Synthèses théoriques	Rédiger une synthèse écrite de 4 pages à partir d'une thématique au choix de l'étudiant (en fonction de la thématique de son mémoire de recherche). L'objectif est d'aider l'étudiant à commencer ses lectures et à le familiariser avec l'écriture de la partie théorique de son mémoire (écriture à venir au 2e semestre)
M2	Compte-rendu de lecture	Dossier de présentation d'un article étudié en cours et apports à la réflexion professionnelle. Volume indicatif du dossier entre 12000 et 15 000 signes (4 à 5 pages dactylographiées)
M2	Compte-rendu de lecture	Devoir sur table (durée 2h30) : Rédigez un compte rendu de lecture synthétique à partir de l'article de G. Vigner, « Réduction de l'information et généralisation : aspects cognitifs et linguistiques de l'activité de résumé », Pratiques n° 72 Le résumé de texte, décembre 1991, 33-54.

NOTES

1. http://childes.psy.cmu.edu/.

2. http://www.ccdmd.qc.ca/fr/.

3. On trouvera en annexes un descriptif de la composition du corpus avec le détail sur les genres de textes, niveaux, disciplines et nombre de mots. Les textes sont accessibles sous plusieurs formats (.doc, .pdf, .txt et .xml) et disponibles en ligne : http://lidilem.u-grenoble3.fr/ressources/corpus-du- labo/article/corpus-litteracie-avancee.

4. Frantext et Scientext sont deux bases en ligne, accessibles ici : http://www.frantext.fr/ ; http://scientext.msh-alpes.fr/scientext-site/spip.php?article9.

5. http://www.tei-c.org/index.xml.

6. On peut toutefois citer l'enquête de Libersan *et al.* (2010) qui aboutit à un inventaire brut de 138 genres listés par les enseignants de *College* à l'UQAM (Université de Québec à Montréal) et l'enquête sur les genres universitaires menée en France et en Belgique dans le cadre du projet ANR EUIPM « Écrits universitaires : inventaire, pratiques, modèles », resp. I. Delcambre (voir par ex. Delcambre et Lahanier-Reuter, 2010).

7. Citons notamment TXM http://textometrie.ens-lyon.fr/ mais aussi l'interface Scienquest, pour une interrogation en ligne (Falaise *et al.*, 2011).

8. Un système d'alignement entre les versions successives d'un texte, à la manière des alignements entre un texte et ses traductions, a récemment été développé comme instrument de recherche en génétique textuelle (-Lardilleux *et al.*, 2013).

9. Des projets de ce type sont en cours en Norvège dans le cadre de l' « Oslo learner corpus of written English and French ») et en Suède dans le cadre du « Corpus écrit de FLE » http://projekt.ht.lu.se/cefle/.

10. La critique génétique discute la question de savoir en quoi ces deux types d'interventions ne sont pas toujours clairement distinctes.
11. Voir l'expérience « Annodis » (Ho-Dac *et al.*, 2014).
12. En ce sens, même si nous en sommes les artisans concrets, le corpus est pour son existence redevable à toute une communauté de recherche.
13. Par exemple l'utilisation de *car* en lien avec la maitrise de l'argumentation.

RÉSUMÉS

Le corpus de littéracie avancée réunit des écrits universitaires et professionnels produits par des étudiants du niveau Licence 1 au Master 2. Il contient actuellement 338 textes (+ d'1 million de mots) et est mis à disposition au format xml, assorti de métadonnées (niveau, discipline, genre, consigne d'écriture etc.). Il est à la fois un aboutissement et un point de départ dans le champ de la littéracie avancée : parce que l'enjeu n'est pas tant de constituer des corpus que de les savoir utilisés et de nourrir les recherches actuelles et futures, il est pensé comme un construit évolutif.
L'article expose d'abord les objectifs du corpus : le premier, d'ordre descriptif est de cerner les compétences rédactionnelles et les besoins de formation à un niveau avancé ; le second, d'ordre didactique, est qu'il sert de ressource pour la formation, notamment pour des activités d'observation et de manipulation par les apprenants. Sont ensuite présentées les métadonnées, conçues en phase avec les recommandations de la TEI et dans la perspective d'un outillage des linguistes et des didacticiens dans le champ de la littéracie avancée. Enfin, l'article met en évidence les potentialités du corpus : intégrer des données sur les processus d'écriture et des textes annotés par des relecteurs-correcteurs, élaborer des dispositifs didactiques à l'échelle textuelle. À terme, le corpus gagnera à être associé à des outils de requête et à s'enrichir d'annotations diverses, selon les besoins des chercheurs et des formateurs.

A corpus of "advanced literacy": result and starting point
The corpus of "advanced literacy" brings together academic and professional texts written by undergraduate and Master's degree students. It currently contains 338 texts (more than 1 million words) and is available in xml format with meta-data (level, discipline, text genre, writing instructions etc.). It is both an outcome and a starting point in the field of advanced literacy. Besides the corpus building, the challenge is that it is used and useful for current and future research projects. Consequently, it is designed as a construction in process.
The present article first displays the objectives of the corpus: the first is to identify the writing skills and training needs at an advanced level; the second is to teach with the corpus as a resource for learners: texts and descriptions serve to develop didactic activities based on observation and manipulation. Then the article comes to the metadata designed according to the recommendations of the TEI and so as to provide linguists and didacticians with an effective tool to achieve their goals. Finally, the article highlights the opportunities of such a corpus: integration of new data about the writing process (drafts, real-time and online writing processes), compilation of texts annotated by teachers-proofreaders, and development of didactic tools on textual skills such as coherence. Eventually, the corpus will be even more efficient when it is associated with query tools and will gradually be enriched with various annotations, *according to the needs of researchers and trainers.*

INDEX

Keywords : literacy, written production, didactic corpus, metadata, high-level annotations
Mots-clés : littéracie, production écrite, corpus didactique, métadonnées, annotations de haut-niveau

AUTEURS

MARIE-PAULE JACQUES

ESPE de Grenoble et Laboratoire Lidilem - Université Grenoble Alpes

FANNY RINCK

ESPE de Grenoble et Laboratoire Lidilem - Université Grenoble Alpes