CORPUS

Corpus

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

Constitution d'un grand corpus d'écrits émergents et novices : principes et méthodes

Sarah De Vogüé, Natacha Espinoza, Brigitte Garcia, Marie Perini, Frédérique Sitri et Marzena Watorek

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/2737
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
Sarah De Vogüé, Natacha Espinoza, Brigitte Garcia, Marie Perini, Frédérique Sitri et Marzena Watorek, « Constitution d'un grand corpus d'écrits émergents
et novices : principes et méthodes », *Corpus* [En ligne], 16 | 2017, mis en ligne le 06 janvier 2018, consulté le 09 janvier 2018. URL : http://journals.openedition.org/corpus/2737

Constitution d'un grand corpus d'écrits émergents et novices : principes et méthodes

Sarah De Vogüé, Natacha Espinoza, Brigitte Garcia, Marie Perini, Frédérique Sitri et Marzena Watorek

1. Introduction

1 La recherche présentée ici vise, à terme, à dégager les invariants et les spécificités dans la construction et le développement de la littératie sur la base d'un vaste corpus de textes écrits recueilli auprès de publics diversifiés. La notion de littératie, issue des recherches de Goody en anthropologie, désigne un champ de recherche interdisciplinaire (psychologie, sociologie, anthropologie, sciences du langage, didactique) dont le point commun est, pour aller vite, de reposer sur une vision dynamique de l'écrit et de prendre en compte « la diversité et [...] la complexité croissante des tâches et des pratiques liées à la lecture et à l'écriture » (Barré de Miniac 2003 : 116). Les notions de dynamique et de diversité sont bien au cœur de notre projet : nous concevons la littératie comme relevant de *processus dynamiques* d'acquisition et d'acculturation à divers genres et types d'écrits, que nous souhaitons saisir à différents moments de leur émergence. Ainsi, l'enjeu central de *Dynascript* [1] est la constitution d'un vaste corpus d'écrits qui permette d'approfondir notre compréhension des processus en jeu dans l'accès à la littératie dans sa diversité, c'est à dire envisagée au travers de la pluralité des genres et types discursifs qui la constituent et chez des apprenants de profils divers, ces derniers croisant ici trois grandes variables : enfants/adultes, langue 1 / langue 2, entendants/sourds.

2 Nous nous focalisons dans cet article sur les principes et méthodes guidant la constitution et la documentation de ce corpus. Ainsi, après avoir brièvement exposé nos options théoriques et nos questions de recherche, puis précisé nos choix en matière de publics d'apprenants, nous nous centrons sur la définition des tâches rédactionnelles demandées

pour, sur ces bases, faire état des questions posées et des options retenues en matière d'outillage de notre corpus.

2. Littératie en construction : perspectives théoriques et questions de recherche

3 Les processus d'accès à la littératie sont ici appréhendés dans une perspective fonctionnaliste et énonciative et au travers de deux ensembles théoriques complémentaires. Du point de vue de l'acquisition des langues, nous nous intéressons à la conception décrite comme « Approche des lectes d'apprenants » (ALA). Un point central de l'ALA (Klein et Perdue 1997, Watorek et Perdue 2005) est de considérer les productions d'apprenants non comme déviance par rapport à *la* norme mais bien comme manifestations (idiolectales) de systèmes linguistiques *per se* dont il s'agit de dégager les normes propres. Ces systèmes, dynamiques et hétérogènes par nature, évoluent graduellement vers la langue cible même s'ils peuvent se fossiliser avant d'atteindre cet état. L'ALA a identifié trois grands stades dans l'acquisition initiale d'une L2 à partir de l'étude des productions orales en L2 d'apprenants de langues cibles diverses et locuteurs de différentes langues sources. Un parcours acquisitionnel complet en six stades a par ailleurs été proposé à partir des productions orales en français d'apprenants suédophones (Bartning et Schlyter, 2004 ou Schlyter, 2003). Ces propositions étant issues de recherches sur des productions orales, il s'agira de les confronter à la problématique de l'acquisition de l'écrit, travail amorcé notamment par Hellqvist (2010) et Tahery (2012).

4 C'est pareillement selon une conception dynamique que nous envisageons la pluralité des genres et types discursifs que recouvre la littératie, le genre étant lui-même conçu comme une catégorie souple (Mellet et Sitri 2010) et non comme un ensemble de normes figées (Rinck et Sitri 2012). Nous nous inscrivons à cet égard en filiation des travaux, d'inspiration bakhtinienne, qui considèrent l'hétérogénéité comme constitutive de toute production verbale. Toute production écrite est considérée comme articulant des logiques discursives multiples et hétérogènes tant sur le plan énonciatif que sur les plans syntaxique et rhétorique. Ces logiques correspondent à des types et à des genres différenciés, à des fonctions et à des pratiques sociales également diverses. Ainsi, pour un scripteur, il s'agit de, tout à la fois, narrer et énumérer, rendre compte et recommander, intéresser et argumenter, par exemple. Dans cette perspective, le regard sur les dynamiques acquisitionnelles se centre sur l'interaction entre les compétences « micro-linguistiques » (orthographe, syntaxe, morphologie) et les caractéristiques des genres et des types dont relève le texte produit.

5 Nos questions de recherche sont de trois ordres. Faisant fond sur les résultats de l'ALA en matière de productions de type oral, un premier enjeu à terme est de dégager des itinéraires d'acquisition éventuellement spécifiques à l'écrit. De manière plus pointue, nous visons à identifier de potentiels invariants dans les processus d'émergence de l'écrit et/ou d'acculturation à de nouveaux genres et types d'écrit en transversal de la diversité des profils d'apprenants et des genres et types discursifs. À titre d'exemple, des travaux ont déjà montré, pour l'oral, une proximité entre les systèmes initiaux en L2 et les systèmes des enfants en L1 (Basic Childe Grammar, Slobin 1985) : cette proximité se retrouve-t-elle à l'écrit et dans quelle mesure dépend-elle, le cas échéant, des genres/types discursifs ? Trouve-t-on la même proximité avec des enfants L2, des enfants

sourds ? À l'inverse, un autre objectif est d'identifier, toujours au regard des profils et des genres, la variation et les spécificités probables des parcours acquisitionnels. Quelles variables interviennent et comment pour spécifier ces parcours ? Enfin, nous espérons au travers de ces observations avancer dans la compréhension de ce que recouvre l'écrit en général et, en particulier, la notion de texte écrit. Les caractéristiques de l'écrit nous paraissent en effet devoir être liées aux modalités spécifiques de son acquisition et aux propriétés des textes.

6 Pour tenter de répondre à ces questions, nous avons décidé de travailler à partir d'un corpus constitué de façon expérimentale croisant la diversité des profils de scripteurs et la diversité des types de textes produits.

3. Options méthodologiques pour la constitution du corpus : profils et tâches

7 Les spécificités du corpus que nous voulons constituer nous ont conduites à accorder une attention particulière aux questions méthodologiques.

3.1 Le choix des publics

8 Pour ce qui est du choix des profils, nous avons fait intervenir des variables portant d'une part sur la variation langue maternelle (LM) / langue 2 (L2), d'autre part sur l'âge et le moment dans le processus d'acculturation à l'écrit.

9 Pour ce qui concerne la variation LM/L2, notre projet présente la spécificité d'intégrer en outre un public sourd, pour qui se pose la question du statut respectif de la langue française écrite et de la langue des signes française (LSF). Les parcours linguistiques des sourds, dont 95 % sont issus de parents entendants non signeurs, sont très diversifiés à cet égard et il est souvent complexe de déterminer clairement quelle est leur langue « maternelle », ou première. Des travaux comparant des textes de sourds et d'entendants ont toutefois montré qu'une partie importante des régularités observées dans les textes de sourds sont communes à celles observées dans les textes des entendants de français langue seconde (Lacerte 1989, Nadeau et Machabée 1998, Perini 2013), tandis que d'autres régularités leur semblent spécifiques. Ceci a pu permettre de poser que l'écrit de la langue vocale est à considérer, pour ces publics, comme une L2. Par ailleurs, compte tenu des particularités que l'on observe dans ces textes de sourds, on peut postuler l'existence de modalités de traitement de l'écrit qui sont spécifiques aux sourds (Garcia et Perini 2010, Perini 2013). Ces deux conclusions, tirées de corpus peu étendus et portant quasi exclusivement sur des récits, méritent toutefois d'être corroborées et affinées.

10 Par ailleurs, pour saisir des processus liés à l'émergence et à la construction de la littératie dans sa diversité chez différents profils d'apprenants, nous avons opté pour ce qui nous apparaît comme des moments clés d'acculturation à de nouveaux genres de l'écrit, et notamment à l'élaboration de textes construits, adressés à un destinataire absent, impliquant une certaine distance (voir Koch et Oesterreicher 2001). Nous avons ainsi constitué trois groupes de 160 sujets chacun (G1, 2, 3) correspondant à ces étapes et au sein desquels se déclinent quatre profils de scripteurs : les entendants français langue maternelle (FLM), les entendants français langue étrangère ou seconde (FLES), les sourds

dont la langue de référence est la LSF et les sourds dont la langue de référence est le français.

11 **G1 :** enfants scolarisés, confrontés à la fois à l'émergence de l'écrit et à l'acculturation aux genres scolaires[2]. Nous nous centrons sur l'étape où la production de textes devient possible pour ces enfants, ce qui suppose des ajustements en fonction des quatre profils définis ci-dessus, notamment en terme d'âge, pour que la tâche puisse être réalisée par tous pour des conditions de production comparables (7-9 ans pour les FLM, 9-11 ans pour les sourds).

12 **G2 :** adultes ayant un degré de littératie évalué approximativement à 2 (sur les 5 niveaux de capacité définis par l'OCDE[3]) qui doivent produire des genres professionnels structurés [4]. On pourra avoir recours à des centres de formation dispensant des actions de remédiation, ce qui facilitera la tâche des chercheurs, chaque apprenant ayant fait l'objet d'un test de positionnement.

13 **G3 :** étudiants ayant un degré 4 de littératie (ou passage 3 à 4) mais qui sont face à de nouveaux genres (demandés dans la sphère universitaire), situation qui peut être source de difficultés voire de dysfonctionnements[5].

14 Nous pourrons ainsi comparer les productions en fonction des différents paramètres retenus : enfants/adultes ; faiblement/fortement littéraciés ; sourds/entendants ; FLM/FLES. La diversité des publics et la dimension exploratoire de notre corpus nous conduisent à apporter un soin particulier à la définition des tâches demandées et à leur mode de passation, de façon à prendre en compte les spécificités de chacun : rapport particulier des sourds à l'écrit (voir notamment Garcia 2010, Perini 2013), moindre maturité cognitive pour les enfants, formulation de la consigne pour les sourds et pour les L2.

3.2 Les tâches

15 La définition des tâches répond à différents impératifs.

16 a) Étant donné la diversité et la forte hétérogénéité des genres et types de textes que les scripteurs sont amenés à produire, nous faisons l'hypothèse qu'écrire implique la mobilisation de nombreuses logiques d'écriture hétérogènes. Nous avons donc choisi de travailler sur plusieurs types ou genres de textes à travers plusieurs tâches d'écriture.

17 La question des typologies des textes et/ou des discours est abondamment discutée dans la littérature avec, d'une part, des problèmes terminologiques (type ou genre ou séquence ou mode ou énonciation ou dominance ou cadres interprétatifs ou plans de texte etc.), d'autre part et surtout, des débats et des interrogations sur les différents niveaux de classement (textes entiers, ou parties de textes), sur les principes de classement (types de marques récurrentes, opérations de repérage énonciatif, composition des structures séquentielles et plans de textes, etc.[6]), et sur la nature même du classement (en catégories étanches, en prototypes, en ressemblances de famille, en catégories universelles, en catégories historiques et évolutives, etc. ; voir Bronckart 2008 pour une discussion). Sans préjuger de la bonne typologie, nous avons opté pour une palette de tâches qui ne parcourent pas tous les possibles, mais permettent de mettre en œuvre les différents paramètres impliqués dans ces classements. Sont par conséquent demandés aux publics de notre échantillon de produire du narratif, du descriptif, de l'argumentatif, de l'expositif[7], qui renvoient à différents types de composition.

18 A été ajoutée une autre forme de séquentialité, plus propre à l'écrit, celle des listes. Quelle que soit sa nature propre, l'écrit mobilise on le sait une autre dimension que l'oral, à savoir la dimension spatiale, dimension que la liste met notamment en œuvre. Par ailleurs, si l'on suit Goody, l'écrit mobilise des formes particulières de savoir, attachées d'une part aux propriétés de stockage et de matérialisation associées à ses supports, au travail de classement et de catégorisation que cette matérialisation apporte d'autre part[8].

19 Par ailleurs, nous avons aussi voulu faire varier les modalités de repérage énonciatif. Elles sont impliquées déjà dans les différences entre narration, exposition et argumentation, la narration relevant de la posture énonciative décrite par Benveniste comme étant celle de l'Histoire, et l'exposition relevant de ce que Hamburger puis Bronckart ont décrit en termes de discours théorique, tandis que l'argumentation implique un discours plus situé, au moins dans sa dimension persuasive où l'interaction avec l'interlocuteur est prégnante.

20 Manquaient cependant des modes de repérage importants, que nous avons intégrés en distinguant entre la restitution d'un récit de fiction et l'élaboration d'un récit personnel : l'opposition fiction/personnel renvoie à un jeu de distance, tandis que l'opposition restitution/élaboration renvoie plutôt à l'origine énonciative de la construction du récit. Nous émettons l'hypothèse que ces récits donneront des variations importantes pour ce qui concerne les embrayeurs et pour ce qui concerne les temps verbaux.

21 Par ailleurs, nous avons choisi d'intégrer des formes correspondant à des répliques de dialogues, mobilisant des actes de langage qui impliquent locuteur et interlocuteur : d'une part la demande, d'autre part la recommandation, correspondant à deux actes directifs mais à des relations hiérarchiques inversées avec l'interlocuteur. Pour ne pas multiplier davantage les tâches, ces paramètres ont été croisés avec les précédents : c'est l'argumentation qui a pour finalité une demande ; et c'est la liste qui est liste de recommandations (voir la liste des six tâches en annexe). Nous espérons ainsi avoir un panel suffisamment représentatif des paramètres impliqués dans la construction d'écrits[9].

22 b) Une des difficultés était d'arriver à ce que les tâches proposées puissent être réalisées par nos différents publics. Pour atteindre cet objectif, il est essentiel de définir des consignes efficaces et de choisir des supports déclencheurs propres à contextualiser les tâches à réaliser de manière crédible. Pour cette raison, il a été admis que pour certaines tâches le genre de textes attendu puisse varier en fonction du public, cela aussi bien sur le plan thématique, sur le plan des situations et des fonctions sociales impliquées, que sur le plan de la complexité des opérations cognitives mobilisées.

23 Lorsque certaines tâches paraissent soit trop complexes, soit simplement trop dénuées d'intérêt pour un type de public, les leur imposer fausserait le résultat. C'est la raison pour laquelle si nous avons conservé, lorsque c'était possible, une consigne, une thématique et un genre commun, il est des cas où les tâches diffèrent selon les publics. Nous avons pris soin de définir des tâches à la fois accessibles et dignes d'intérêt et de prévoir des consignes privilégiant des situations pragmatiques (« vous écrivez à untel… » « faites-moi une liste » « laissez un mot à untel pour lui rappeler…. »), mais aussi des situations suffisamment motivantes et intéressantes[10]. Sans doute les tâches demandées ont-elles toutes un caractère artificiel, les textes à produire étant des textes rédigés pour le compte de ce programme de recherche. Ce caractère artificiel permet cependant d'arriver à construire un corpus constitué d'écrits aussi comparables que possible puisque rédigés dans des conditions équivalentes et selon des modalités comparables. Notre

corpus n'est donc pas constitué au moyen d'un recueil écologique de données mais de manière expérimentale, à partir de questions de recherche et en fonction des problématiques visées.

24 Une attention particulière sera portée à la formulation des consignes, en veillant notamment à préciser systématiquement le destinataire du texte, destinataire qui joue un rôle crucial pour la construction d'un texte élaboré, dans la mesure où il est, en tant que lecteur, la mesure même de la lisibilité de ce texte et de son caractère fini.

25 Pour éviter cependant des effets de connivence et de références implicites communes qui se substitueraient à l'explicitation écrite, pour que la part dialogale et le genre de la lettre adressée ne dominent pas l'ensemble des écrits produits, il nous semble important que le destinataire ne soit pas pour autant dans une relation de familiarité avec le scripteur, mais surtout que le texte qui lui est destiné ne lui soit pas à proprement parler adressé. L'équilibre paraît d'autant plus complexe à obtenir qu'il faut aussi que l'écrit réponde à une motivation effective, qui passe par un intérêt supposé du destinataire en question pour la rédaction de ce texte. Or un destinataire qui manifeste son intérêt pour un texte, devra de ce fait même être dans une relation d'échange avec le scripteur, avec demande d'une part, partage d'intérêts communs d'autre part. Des dispositifs tels qu'un forum internet, un journal d'école ou la situation d'un étranger demandant à être informé peuvent faire entendre ce positionnement complexe.

26 Par exemple, pour la tâche de restitution d'un extrait de fiction, les consignes proposées pourront être : « Vous participez à un forum internet. Un internaute a raconté un extrait de Tom et Jerry que vous n'aviez pas vu et qui vous a intéressé. Vous lui écrivez pour lui raconter à votre tour l'extrait que vous avez vu. Il s'agit de partager vos expériences respectives. »[11] La formulation devra, bien sûr, être adaptée à chaque type de public. Par exemple, pour les enfants, on pourra avoir : « Ton copain a vu un extrait du dessin animé Tom et Jerry mais ce n'est pas le même que celui que tu as vu. Tu vas lui raconter l'extrait que tu as vu. » De même, pour la tâche de description, la consigne sera la suivante : « Un ami/un parent qui est absent et qui ne connaît pas le lieu où vous habitez vous a écrit pour vous demander de lui décrire ce lieu qu'il aimerait pouvoir se représenter. Vous lui répondez en lui décrivant votre logement/maison de manière à ce que votre destinataire puisse se l'imaginer. »

27 Nous prévoyons aussi de consacrer un temps collectif à discuter des expériences des scripteurs concernant les types de textes et les contextes évoqués. Par exemple, pour la description, nous pourrons discuter avec les scripteurs de l'existence d'absents dans leur vie et leur demander d'évoquer leur lieu de vie. Nous pourrons également demander aux scripteurs s'ils ont déjà reçu ou écrit des lettres de ce type : on écrit en effet en relation avec d'autres écrits et il faut en tenir compte. Une synthèse des informations recueillies pour l'ensemble du groupe sera consignée dans les métadonnées.

28 c) Étant donné le mode de constitution et l'ampleur du corpus il ne paraît pas possible d'observer les différentes étapes de la rédaction. Aucune consigne ne sera donc donnée quant aux modalités de la rédaction, qui pourra passer ou non par des étapes de type brouillon, esquisse, reformulation, plan, etc. En revanche, il sera demandé aux scripteurs de nous laisser leurs éventuels brouillons ou pré-textes s'ils y consentent. Ce corpus de brouillons sera conservé et utilisé comme corpus secondaire mobilisable éventuellement pour compléter l'analyse du corpus premier.

29 Un premier descriptif des tâches est donné en annexe. Pour chacune d'elles, il est prévu de réaliser un pilote. Les tâches précises, les consignes et les modalités de passation ont par conséquent vocation à être rediscutées en fonction des résultats obtenus pour ces pilotes.

4. Une exploration automatisée du corpus : pourquoi et comment ?

30 Si nous avons choisi une exploration automatisée du corpus, c'est tout d'abord en raison de la grande taille du corpus que nous envisageons de constituer[12] : les outils informatiques permettent d'exploiter de grandes masses de données, en automatisant les calculs de fréquence et rendent possibles des calculs (spécificités, cooccurrences) impossibles à réaliser « à la main ». En outre, à l'heure actuelle, la plupart des logiciels permettent de traiter non seulement les unités graphiques mais aussi des unités catégorisées en fonction des objectifs de la recherche : catégories morpho-syntaxiques annotées automatiquement ou catégories pragmatiques, sémantiques ou discursives nécessitant une annotation manuelle. La possibilité offerte par certains outils de mettre en relation les unités ainsi catégorisées et même de visualiser les « chaînes » ainsi constituées (voir par exemple dans Landragin *et al.* 2012) apparaît particulièrement adéquate à une analyse textuelle telle que celle que nous comptons mener. Le partitionnement du corpus permet en effet de mettre en relation les calculs effectués sur les unités sélectionnées avec les différentes parties du corpus et offre la possibilité d'une analyse contrastive, qui constitue précisément le cœur de notre projet.

31 Mais il faut souligner également le caractère heuristique d'un tel recours à l'outillage dans le cadre de *Dynascript*. D'une part, en effet, l'un des premiers intérêts des processus d'automatisation est, comme le soulignait Franck Neveu dans son introduction à une journée d'études organisée à Paris 3 autour des corpus scolaires[13], de faire apparaître de nouveaux observables. D'autre part, la réflexion préalable au choix du mode d'exploration des données – choix du logiciel d'exploration et choix des annotations éventuelles – suppose une discussion collective sur les catégories pertinentes à relever en relation avec les objectifs de la recherche. Ces catégories sont aussi issues d'un dialogue – et éventuellement d'un compromis – entre des approches différentes (qu'il s'agisse de l'objet d'étude ou des cadres théoriques).

32 Face à la palette d'outils actuellement disponibles, nous avons décidé de produire d'abord un cahier des charges, à partir du travail de description des tâches et des types de textes attendus.

4.1 Choix préalables à la constitution du cahier des charges

33 Les choix relatifs à la numérisation et à la préparation du corpus sont d'autant plus importants que l'un des objectifs opérationnels de *Dynascript* est de mettre à la disposition de la communauté l'ensemble des corpus constitués, ce qui passe par un format d'encodage des fichiers garantissant la pérennisation et l'inter-opérabilité, comme le format XML-TEI. Étant donné la nature des données, il faut également prévoir le mode de traitement des formes « non normées » en conservant conserver la trace de cette forme tout en donnant la possibilité au logiciel d'opérer des calculs sur la forme normée. Par

ailleurs, en raison de la diversité des publics étudiés, le renseignement des métadonnées doit être particulièrement soigné. Il s'agit ainsi pour chaque texte de documenter une fiche signalétique comportant trois niveaux d'informations : (i) identité du texte (code, date rédaction, nom du transcripteur) ; (ii) données sur le scripteur (nom codé, âge, niveau scolaire, langues connues, catégorie de scripteur, type de tâche, degré de familiarité avec les types de textes attendus...) ; (iii) conditions de production (consigne, lieu de passation, personnes présentes, aide fournies, textes déjà connus du scripteur pouvant lui servir de modèle...).

4.2 Cahier des charges pour un outil d'exploration automatique

34 Pour établir notre cahier des charges, nous sommes parties des caractéristiques que nous avons identifiées *a priori* comme étant celles des types de textes correspondant aux tâches demandées aux scripteurs, l'objectif étant de dégager de façon contrastive les caractéristiques des textes produits selon les publics. On notera que cette évaluation peut mettre en jeu des critères de type qualitatif – adéquation à la consigne, caractère compréhensible du texte, « mise en intrigue » pour les textes narratifs ou équilibre entre émotif et rationnel pour la demande argumentée.

35 Si l'on s'en tient aux traits évaluables quantitativement, il faut tout d'abord prendre en compte, dans l'analyse, les caractéristiques « transversales » aux tâches demandées et aux types de textes attendus : longueur, richesse du vocabulaire, complexité syntaxique, nombre de paragraphes, nombre de signes de ponctuation, proportion de signes d'émotion. On peut ensuite dégager pour chaque type de texte, à partir du descriptif des tâches présenté dans la section précédente, une liste « d'observables » mesurables quantitativement. Ainsi pour la tâche 1 présentée dans la section 2.

Temps verbaux
Procès (processus, état, accomplissement, etc.)
Connecteurs temporels
Entités nommées : protagonistes, temps, espace
Chaînes référentielles
Type de progression
séquences descriptives, explicatives, argumentatives
Plan de texte
Premier plan/arrière plan

Tableau 1. observables de la tâche 1

36 En partant du type d'items que nous souhaitons observer et des types d'exploration que nous voulons mener, nous définissons les fonctionnalités d'un outil d'exploration

automatique (d'un logiciel). Nous voulons ainsi pouvoir compter des formes graphiques et mettre en relation le nombre d'occurrences avec le nombre total de formes afin de mesurer la richesse du vocabulaire. Nous devons aussi pouvoir compter des catégories morphosyntaxiques (verbes, noms, adjectifs, adverbes, déterminants, connecteurs, auxiliaires, pronoms personnels), ce qui suppose une annotation morphosyntaxique. Les calculs ainsi que le type d'annotation que nous souhaitons effectuer sont opérés par des analyseurs automatiques intégrés à la plupart des logiciels de textométrie (comme Tree Tagger, dans *le Trameur* ou TXM) moyennant des adaptations ou corrections manuelles. De même, il nous faut pouvoir compter des catégories sémantiques (entités nommées, types de procès, valeurs temporelles et aspectuelles des marques temporelles, valeurs des déterminants, marques d'appréciation, marques d'émotion, expressions temporelles, termes d'adresse...) ce qui suppose de les annoter manuellement. Nous avons également besoin de compter des segments de texte (premier plan/arrière plan, propositions principales/propositions subordonnées, expressions référentielles, segments thématiques/segments rhématiques, séquences narratives/descriptive/dialogales/ argumentatives, énoncés définitoires...) ce qui suppose là encore la possibilité soit de marquer ces segments par des jalons textuels, soit de les annoter manuellement[14].

37 Finalement, nous voulons pouvoir mettre en relation certains segments de texte afin de constituer des chaînes : chaînes référentielles, types de progression par exemple, expressions de premier plan *vs* arrière plan. Il est alors nécessaire de prévoir un outil *ad hoc* : il peut s'agir des logiciels développés pour l'annotation dite « discursive » (type Glozz, Annodis ou Analec pour les chaînes de référence) ou de logiciels de textométrie offrant des fonctionnalités d'annotation tels que le Trameur (voir Fleury et Zimina 2014). Une question sera de savoir s'il est préférable de travailler avec deux outils complémentaires au risque de ne pas pouvoir croiser les résultats de l'un et de l'autre ou plutôt d'utiliser un seul outil même s'il est moins performant sur certaines fonctionnalités.

38 La recension des catégories pertinentes, établie ici *a priori*, devra être précisée grâce à l'interaction entre observation des premiers textes et fonctionnalités de ou des outils sélectionnés. Cette phase d'ajustement nous permettra de nous accorder sur des catégorisations dont le découpage peut être différent selon les cadres théoriques : c'est le cas par exemple des notions de « premier plan/arrière-plan » dans le modèle de la Quaestio et dans l'approche de Labov ou de la linguistique textuelle. Nous serons également amenées à prendre des décisions concernant des phénomènes récursifs comme les enchâssements de subordonnées ou de groupes. Ainsi, pour reprendre Habert (2005), « l'annotation suppose en amont un consensus relatif sur la manière de découper et d'étiqueter en fonction des phénomènes visés » (123).

Conclusion

39 Le travail de constitution et d'exploration du corpus s'annonce conséquent. Un tel corpus cependant, prenant en compte des profils différents, des moments différents dans le processus d'acculturation à l'écrit et des types de textes différents, répond au besoin qui est le nôtre de disposer d'un état élaboré des invariants et des variations relatifs à l'émergence de l'écrit en tant que travail de production de textes : des textes pris dans leur diversité et dans ce qui peut constituer leur difficulté pour chacun.

40 Cette phase d'élaboration des données constitue de fait une étape cruciale de notre recherche. Nombre de nos questionnements – métadonnées, traitement des formes « déviantes » par exemple – sont sans nul doute partagés par les concepteurs de corpus de scripteurs en phase d'apprentissage. Divers éléments de méthode sont en outre transférables à toute recherche linguistique basée sur des corpus, et notamment la réflexion sur un cahier des charges en préalable au choix d'un outil d'exploration automatisée.

41 Plus généralement, nous voudrions défendre notre option de travailler sur des données non pas recueillies mais produites de manière expérimentale : si l'on perd ainsi le caractère « naturel » de productions produites dans un environnement écologique (l'école, l'université, le travail), tout le travail en amont pour penser la constitution du corpus et son exploration permet de construire des observables inscrits dans une problématique – celle de l'émergence et de la variation de l'écrit. Culioli soutient qu'il ne saurait y avoir « d'observables sans théorie des observables » : « les observations ne sont pas données mais construites par une démarche qui (...) multipli(e) l'empirique de façon controlée » (Culioli 1990, p. 19). C'est là une ligne relativement peu développée dans la linguistique de corpus. Et dans les faits, il n'est pas question de la suivre à la lettre, puisqu'il ne sera pas question de faire proliférer les données comme le fait Culioli en permutant, substituant, manipulant. En revanche, il est bien question de raisonner sur ces données, et les catégories que nous mettrons au point pour les annotations comme pour les métadonnées seront de ce point de vue décisives non pas pour aboutir à un morcellement ou à un cloisonnement mais pour comprendre problèmes et dynamiques qui constituent l'écrit dans son hétérogénéité et dans son émergence, c'est-à-dire pour mieux comprendre les pro-cessus en jeu dans l'accès à la littératie sous toutes ses formes.

BIBLIOGRAPHIE

Adam J.-M. (1990). *Éléments de linguistique textuelle*. Liège : Mardaga.

Adam J.-M. (1999). *Linguistique textuelle. Des genres de discours aux textes.* Paris : Nathan.

Bakhtine M. (1984) [1952-1953]. « Les genres du discours », *in* : *Esthétique de la création verbale*. Paris : Gallimard.

Baroni R. (2007). *La Tension narrative*. Paris : Seuil.

Barré-de Miniac C. (2003). « La littéracie : au-delà du mot, une notion qui ouvre un champ de recherches variées », *Revue suisse de sciences de l'éducation* 25 : 111-123.

Bartning I., Schlyter S. (2004). « Itinéraires acquisitionnels et stades de développement en français L2 », *Journal of French Language Studies*, 14 : 281-299.

Benazzo S. (2012). « Language origins, Learner Varieties and Creating Language Anew : How Acquisitionnal Studies Can Contribute to Language Evolution Research », *in* M. Watorek, S. Benazzo, S. Hickmann (éd.) *Comparative Perspectives on Language Acquisition : Tribute to Clive Perdue.* Bristol : Multilingual Matters : 204-222.

Benveniste E. (1966). « Les relations de temps dans le verbe français », *in* : *Problèmes de linguistique générale*, t. I. Paris : Gallimard : 237-250.

Bronckart J.-P. (2008). « Genres de textes, types de discours et "degrés" de langue », *Texto !* XIII, 1.

Combettes B. et Tomassone R. (1988). *Le texte informatif, aspects linguistiques*. Bruxelles : De Boeck-Wesmael.

Culioli A. (1990). *Pour une linguistique de l'énonciation*, tome 1. Paris : Ophrys.

De Vogüé S., Lehmann S., Sitri F. (2014). *Atelier de Langue Française*. UPOND-Comète.

De Vogüé S. (2014). « Effets sémantiques, syntaxiques et énonciatifs du jeu entre quantité et qualité », *LINX* 70-71 : 141-164.

Espinosa N., Vertalier M., Canut E. (2014). *Linguistique de l'acquisition du langage oral et écrit*. Paris : L'Harmattan.

Fleury S., Zimina M. (2014). « Trameur : A Framework for Annotated Text Corpora Exploration », *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics : System Demonstrations*. Dublin : 57-61.

Fusellier-Souza I. (2012). « Multiple Perspectives on the Emergence and Development of Human Language : B. Comrie, C. Perdue and D. Slobin », *in* M. Watorek, S. Benazzo, S. Hickmann (éd.) *Comparative Perspectives on Language Acquisition : Tribute to Clive Perdue*. Bristol : Multilingual Matters : 204-222.

Garcia B., Perini M. (2010). « Normes en jeu et jeu des normes dans les deux langues en présence chez les sourds locuteurs de la Langue des Signes Française (LSF) », *in* B. Garcia, M. Derycke (coord.), *Sourds et langue des signes. Norme et variations*, Numéro spécial de la revue *Langage et Société* 131 : 75-94.

Garnier S., Rinck F., Sitri F., de Vogüé S. (à paraître). « Former à l'écrit universitaire : un terrain pour la linguistique », *Linx* 72.

Goldin-Meadow S., Mayberry R.I. (2001). « How do Profoundly Deaf Children Learn to Read ? », in *Learning Disabilities Research and practice* 16 : 221-228.

Goody J. (1979 pour la traduction). *La Raison graphique. La domestication de la pensée sauvage*. Paris : Éditions de Minuit.

Habert B. (2005). *Instruments et ressources électroniques pour le français*. Paris : Ophrys.

Hamburger K. (1977, 1986 pour la traduction). *Logique des genres littéraires*. Paris : Seuil.

Hellqvist B. (2010). « La subordination dans la production écrite en L2. Étude sur l'acquisition de la subordination en français L2 chez 10 lycéens suédophones ». Mémoire de master, Linnaeus University.

Klein W., Perdue C. (1997). « The Basic Variety. Or : Couldn't Natural Languages be much Simpler? », *Second Language Research* 13 : 301-347.

Koch P., Oesterreicher W. (2001). « Langage oral et langage écrit », in G. Holthus (éd.) *Lexicon der Romanistischen Linguistik* , tome 1-2. Tübingen : Max Niemeyer : 584-627.

Lacerte L. (1989). « L'écriture sourde québécoise », *Revue québécoise de linguistique théorique et appliquée* 8, 3-4 : 303-345.

Labov W. (1972). « The transformation of Experience in Narrative Syntax », *in* : *Language in the inner city*. University of Pennsylvania Press : 354-396.

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

Lenart E. (2011). « La compétence discursive en langue étrangère dans l'enseignement pré-secondaire », *in* G. Komur, P. Trevisiol (éd.) *Discours, didactique et acquisition des langues : les termes d'un dialogue*. Orizons : 363-377.

Mellet C., Sitri F. (2010). « Nom de genre et institutionnalisation d'une pratique discursive : le cas de l'interpellation parlementaire et du signalement d'enfant en danger », *Congrès Mondial de Linguistique Française, La Nouvelle Orléans*. http://dx.doi.org/10.1051/cmlf/2010175.

Nadeau M., Machabee D. (1998). « Dans quelle mesure les erreurs des sourds sont-elles comparables à celles des entendants ? », *in* C. Dubuisson et D. Daigle (dir.) *Lecture, écriture et surdité*. Montréal : Les Éditions logiques : 169-195.

OCDE (2000). « La littératie à l'ère de l'information, rapport final de l'enquête internationale sur la littératie des adultes ». http://www.oecd.org/fr/edu/innovation-education/39438013.pdf.

Perdue C. (éd.) (1993). *Adult Language Acquisition : Cross-linguistic Perspectives*. 2 volumes. Cambridge : Cambridge University Press.

Perini M. (2013). *Que peuvent nous apprendre les productions écrites des sourds ? Analyse de lectes écrits de personnes sourdes pour une contribution à la didactique du français écrit en formation d'adultes*. Thèse de Doctorat. Université Paris 8.

Rastier F. (2001). *Arts et sciences du texte*. Paris : PUF.

Reuter Y. (2006). « À propos des usages de Goody en didactique », *Pratiques* 131-132 : 131-154.

Rinck F. et Sitri F. (2012). « Pour une formation linguistique aux écrits professionnels », *Pratiques* 153-154, I. Delcambre, D. Lahanier-Reuter (éd.) *Littéracies universitaires : nouvelles perspectives* : 71-84.

Schlyter S. (2003). « Stades de développement en français L2. Exemples d'apprenants suédophones, guidés et non-guidés, du "Corpus Lund" », Working paper.

Slobin D.I. (1985). « Crosslinguistic evidence for the language making capacity », *in* D.I. Slobin (éd.) *The crosslinguistic study of language acquisition*, vol. 2. Hillsdale, N.J. : Lawrence Erlbaum : 1157-1256.

Tahery Z. (2012). *L'acquisition de la temporalité en français par des apprenants persanophones*, Thèse de doctorat. Université Paris 3.

Vertalier M., Espinosa N. (2010). « Préparer l'enfant non encore lecteur à lire-écrire : quelles interactions avec l'écrit ? », *Caractères* 32 : 34-40.

Watorek M. et Perdue C. (2005). « Psycholinguistic Studies on the Acquisition of French as a Second Language : The 'Learner Variety Approch' », *in* J.-M. Dewaele (éd.) *Focus on French, Multilingual Matters*. Bristol : Multilingual Matters : 1-16.

Watorek M. (éd.) (2004). *La construction du discours en français langue cible, Langages* 155.

ANNEXES

Annexe : Liste des tâches

TÂCHE 1 : Restitution d'un extrait de fiction (film ou dessin animé)

Stimulus : *Tom et Jerry (*Déjà utilisé avec des enfants et adultes sourds dans le cadre du corpus LS-COLIN.) / *Mister Bean* ; extrait de une minute.

Proposition de consigne :

Proposition de consigne : « Vous participez à un forum internet. Un internaute a raconté un extrait de Tom et Jerry / Mister Bean que vous n'aviez pas vu et qui vous a intéressé. Vous lui écrivez pour lui raconter à votre tour l'extrait que vous avez vu. Il s'agit de partager vos expériences respectives. »

Pour l'enfant : « Ton copain a vu un extrait du dessin animé Tom et Jerry mais ce n'est pas le même que celui que tu as vu. Tu vas lui raconter l'extrait que tu as vu. »

TÂCHE 2 : Récit d'expérience personnelle

Proposition de consigne :

« Le journal de votre école / quartier / entreprise / association voudrait publier le récit de votre première arrivée dans l'école / le quartier / l'entreprise / l'association. »

Pour les enfants :

« Tu vas raconter ta première journée dans l'école. Imagine que ton récit sera publié dans le journal de ton école »

TÂCHE 3 : Description de lieu

Proposition de consigne :

« Un ami/un parent qui est absent et qui ne connaît pas le lieu où vous habitez vous a écrit pour vous demander de lui décrire ce lieu qu'il aimerait pouvoir se représenter. Vous lui répondez en lui décrivant votre logement/maison de manière à ce que votre destinataire puisse se l'imaginer. »

TÂCHE 4 : Exposition

Proposition de consigne :

« Un extraterrestre vous a contacté par internet. Il vous interroge sur tout ce qui constitue notre univers. Vous lui avez parlé des moyens de transport. Il vous demande ce qu'est une voiture. Vous lui expliquez. »

TÂCHE 5 : Argumentation + demande

Proposition de consigne :

« Vous écrivez à un responsable pour lui demander une faveur (entretien d'embauche, octroi de points supplémentaires à l'école, octroi de dispense d'assiduité, audience, candidature à une formation, etc.) »

Enfant : « Vous écrivez au directeur de l'école pour lui demander d'organiser une sortie scolaire qui vous plaît »

TÂCHE 6 : Liste de recommandations

Proposition de consigne :

« Un appareil (une cafetière / un aspirateur / un lecteur de DVD) est mis à disposition de tous (amis/collègues/co-locataires/ famille / classe/bureau / passagers) dans un lieu commun (salle de classe, salle de détente, cuisine commune, entrepôt commun, voiture

partagée). Vous êtes chargé-e de rédiger une liste de recommandations pour le bon usage de cet appareil (utilisation, précautions ET rangement), recommandations qui seront affichées à côté de l'appareil (sous la forme d'un post-it ou d'une grande affiche selon le lieu concerné) »

NOTES

1. Le projet Dynascript est financé dans le cadre de l'appel à Projets 2015 de l'Université Paris-Lumières (voir : http://www.u-plum.fr/app/webroot/upload/files/Appel%20à%20projets%202015-2.xlsx.pdf).

2. Voir Espinosa, Vertalier et Canut (2014).

3. Voir ce rapport : OCDE (2000) « la littératie à l'ère de l'information ».

4. Voir aussi Reuter (2006) pour une discussion de ces relations entre savoir et écrit, et des implications pour la didactique à l'école.

5. Voir Garnier *et al.* (à paraitre), pour un ensemble d'études sur les relations entre écrit et savoir à l'université. Les difficultés avec l'écrit font l'objet d'un travail de remédiation élaboré à UPOND (voir de Vogué, Lehmann et Sitri, 2014).

6. Sur ces différents principes de classement, voir bien sûr Benveniste (1966), Adam (1990), Adam (1999), et Bronckart (2008) pour une discussion mettant ces principes en perspective. Dans le cadre de la théorie des opérations énonciatives et prédicatives de Culioli, voir de Vogué (2014).

7. Le choix d'un texte expositif plutôt qu'explicatif pourrait être discuté (voir les débats sur la différence et l'importance relative des deux types notamment dans Combettes & Tomassone 1988) : nous avons supposé qu'un texte expositif pouvait intégrer de l'explicatif et valait plus généralement pour intégrer toute démarche visant à définir et caractériser de manière générale des objets, sans s'inscrire pour autant dans une démarche argumentative (une définition a d'autres propriétés linguistiques qu'une hypothèse, la première faisant partie du discours expositif, la seconde du discours argumentatif).

8. C'est au demeurant une autre raison pour retenir le type de l'exposition qui paraît caractéristique d'un rapport au savoir (définition, élaboration d'objets de savoir) décrit par Goody comme caractéristique de l'écrit.

9. À noter que ne sont pas parcourues en revanche les variations correspondant à ce que Bakhtine a appris à distinguer en termes de genres de discours : des « types relativement stables d'énoncés développés par des sphères sociales » (Bakhtine 1984). Sans doute est-il attendu d'un scripteur, comme de tout locuteur au demeurant, qu'il maîtrise les compétences qui lui permettent, ou permettront dans le cas des enfants, de produire des textes qui s'inscrivent dans ces différents genres ou, au moins, se positionnent par rapport à eux. Ces genres cependant sont nombreux, variés, dépendants par définition des conditions sociales de leurs usages, et sans doute fondamentalement ouverts (voir notamment Mellet et Sitri 2010). Il est clair en particulier que les différents publics que nous souhaitons comparer ont vocation à fréquenter des genres différents puisqu'ils relèvent de sphères sociales d'activité différentes avec des pratiques sociales différentes (voir Rinck et Sitri (2012) pour une discussion sur les genres professionnels). C'est la raison pour laquelle nous avons laissé la possibilité que chaque public définisse pour chaque classe un genre socio-discursif en relation avec ses pratiques sociales propres.

10. Voir la partie « évaluative » des récits décrits par Labov, partie dont la fonction est de montrer l'enjeu du récit (de répondre à la question « so what ? » : il s'agit d'éviter les « pointless stories » (Labov 1972, p. 366).

11. Il s'agit d'éviter un récit de type critique de film qui évaluerait la qualité ou viserait à persuader l'autre de voir ou ne pas voir ce film. Il s'agit donc de minimiser autant que possible la dimension argumentative du récit.

12. Nous prévoyons de collecter au total 2880 textes (6 tâches x 480 sujets).

13. « Analyser informatiquement des grands corpus d'écrits scolaires : problèmes de transcription, d'annotation et de traitement », journée organisée à Paris 3 le 18 mars 2015 par le groupe Écriture Scolaire (EA 7345 CLESTHIA).

14. Le groupe 8 du consortium Corpus Écrit a travaillé sur les annotations de haut niveau. Le wiki du groupe fournit un certain nombre de ressources didactiques et bibliographiques (https://groupes.renater.fr/wiki/corpus-ecrits/public/groupe-8).

RÉSUMÉS

Constitution d'un grand corpus d'écrits émergents et novices : principes et méthodes

Cette contribution propose une réflexion sur la construction d'un vaste corpus d'écrits qui permet d'approfondir notre compréhension des processus en jeu dans l'accès à la littératie dans sa diversité, au travers de la pluralité des genres et des types discursifs qui la constituent et chez des apprenants de profils divers : enfants/adultes, langue 1 / langue 2, entendants/sourds.

La réflexion sur la mise en place de ce corpus s'inscrit dans une perspective fonctionnaliste et énonciative et fait fond sur deux ensembles théoriques complémentaires. Du point de vue de l'acquisition des langues, nous souscrivons à la conception selon laquelle les productions d'apprenants ne constituent pas une déviance par rapport à la norme mais sont des manifestations (idiolectales) de systèmes linguistiques per se dont il s'agit de dégager les normes propres.

Du point de vue de la littératie, nous nous inscrivons dans la tradition des travaux, d'inspiration bakhtinienne, qui considèrent l'hétérogénéité comme constitutive de toute production verbale.

Ainsi, nous expliquons et justifions nos choix en matière de profils d'apprenants et de tâches rédactionnelles, en relation avec la perspective théorique adoptée, ainsi que les options retenues en matière d'outillage du corpus.

Creating a big corpus of emerging written texts: guidelines and methodology

This contribution offers a reflection on the collection of a vast corpus of writings in order to deepen our understanding of the processes involved in access to literacy in its diversity. Indeed, this corpus is very diverse in its essence: it contains a plurality of genres and discourse types, and was made from the productions of learners with varying characteristics: children/adults, language 1 / language 2, hearing/deaf.

In this reflection, we take a functionalist and enunciative approach with two complementary theories. From the perspective of language acquisition, we consider that the learners' productions aren't a deviation from the norm, but are (idiolectal) manifestations of linguistic systems per se whose specific norms we have to identify.

From the perspective of literacy, we share the tradition of the works inspired by Bakhtin's theory, who considers the heterogeneity as an inherent part of any verbal production.

Thereby, we explain and justify our choices of learners' profiles and writing tasks, in relation to the theoretical perspective adopted and the selected options for the structuring of the corpus.

INDEX

Mots-clés : littératie, acquisition de l'écrit, corpus d'écrits, annotation de corpus écrit, genres discursifs, tâches communicatives, langues secondes, langues vocales vs langues des signes, écrits de sourds
Keywords : literacy, acquisition of written French, corpus of written French, annotation of written French corpus, discursive registers, communicative tasks, second language, spoken Vs. signed languages, Deaf's writings

AUTEURS

SARAH DE VOGÜÉ

Modyco-Université Paris-Ouest Nanterre-CNRS

NATACHA ESPINOZA

Modyco-Université Paris-Ouest Nanterre-CNRS

BRIGITTE GARCIA

SFL-Université Paris 8-CNRS

MARIE PERINI

SFL-Université Paris 8-CNRS

FRÉDÉRIQUE SITRI

Modyco-Université Paris-Ouest Nanterre-CNRS

MARZENA WATOREK

SFL-Université Paris 8-CNRS