CORPUS

Corpus

8 | 2009
Corpus de textes, textes en corpus

Du corpus littéraire au corpus linguistique : dématérialisation, restructuration, lectures rhizomatiques et analyses linguistiques des manuscrits

Thomas Lebarbé

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/1694
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Édition imprimée
Date de publication : 15 novembre 2009
Pagination : 221-239
ISSN : 1638-9808

Référence électronique
Thomas Lebarbé, « Du corpus littéraire au corpus linguistique : dématérialisation, restructuration, lectures rhizomatiques et analyses linguistiques des manuscrits », *Corpus* [En ligne], 8 | 2009, mis en ligne le 01 juillet 2010, consulté le 19 avril 2019. URL : http://journals.openedition.org/corpus/1694

Du corpus littéraire au corpus linguistique : dématérialisation, restructuration, lectures rhizomatiques et analyses linguistiques des manuscrits

Thomas LEBARBÉ
Université Stendhal – Grenoble 3
Laboratoire LIDILEM (EA 609)

Préambule

Le terme « corpus » n'est pas l'apanage d'un domaine de recherche en linguistique et fait l'objet de nombreuses définitions selon les disciplines qui se le sont approprié. Cette ambigüité forte rend le dialogue interdisciplinaire complexe mais peut être riche d'enseignements.

Dans cet article, nous présentons le lien étroit que nous avons construit entre les définitions littéraires et linguistiques du terme. Dans le cadre du projet de valorisation scientifique des manuscrits de Stendhal, la notion de page de manuscrit a été formalisée pour être transcrite et annotée en XML par des spécialistes. L'ensemble des pages transcrites peut alors être restructuré, dématérialisé à l'envi, afin d'offrir aux littéraires comme aux linguistes de nouveaux objets d'étude. Le corpus littéraire devient corpus pour la linguistique.

Le projet « Manuscrits de Stendhal »

Les manuscrits de Stendhal ont été rédigés entre 1802 et 1842 et sont particulièrement fragiles : le papier est friable, certaines écritures au crayon s'effacent à cause du frottement des pages entre elles. Le fonds représente 40.000 pages conservées et récemment numérisées par la Bibliothèque municipale de Grenoble. Contrairement aux manuscrits de Mme Bovary, récemment publiés en ligne[1] par les chercheurs de l'université de Rouen, les manuscrits de Stendhal ne sont pas les notes de

1 http://bovary.univ-rouen.fr

Corpus n°8 « Corpus de textes, textes en corpus »
(2009), 221-239

genèse d'une seule et unique œuvre. En effet, les brouillons des œuvres de Stendhal ont, pour la plupart, été systématiquement détruits par les éditeurs une fois l'ouvrage sous presse.

Les manuscrits de Stendhal sont un ensemble de feuillets et de cahiers rassemblés à la mort de l'auteur par ses proches, puis complétés d'acquisitions pour certaines récentes. Les pages ont à l'origine été organisées en registres selon un critère de taille des feuillets plutôt qu'un critère de cohérence diachronique ou littéraire. Au delà du désordre induit par ce classement, chaque page elle-même peut contenir différentes unités textuelles, rédigées à des dates différentes (notes de relectures), dans des objectifs différents (brouillons littéraires, notes diaristes…).

L'équipe littéraire des « Manuscrits de Stendhal » du laboratoire « Traverses 19-21 » a souhaité mettre en place un outil qui permette d'afficher en ligne la page numérisée en vis-à-vis de sa transcription. Un premier prototype, développé avec un système de gestion de base de données propriétaire, permettait cet affichage, mais réduisait la transcription à une simple mise en forme graphique à l'image de la page originale. Bien que répondant au besoin initial d'affichage, cette solution n'apportait aucun moyen de manipuler l'ensemble documentaire.

Par ailleurs, la transcription ne pouvait s'afficher que d'une seule manière, sans sélection de l'information affichée en fonction du type d'utilisateur, sans reformatage possible. Or, force est de constater que l'analyse littéraire des manuscrits nécessite d'autres représentations de l'objet d'études : strates d'écritures (Meynard 2007), versions parallèles (Spengler 2006). La transcription enrichie est une surcharge d'information pour le grand public alors qu'elle constitue un accompagnement nécessaire au travail du chercheur.

L'équipe Manuscrits de Stendhal s'est donc associée aux informaticiens-linguistes du laboratoire LIDILEM afin de construire un modèle de transcription des pages répondant aux divers besoins de représentation et de transformation des manuscrits. Outre le défi technique d'ingénierie documentaire, l'intérêt linguistique réside dans le fait que les manuscrits du fonds Stendhal correspondent à 40 ans (1801-1842) de pratique

scripturale autographe[2] et allographe[3]. C'est par conséquent autour du manuscrit transcrit, structuré et annoté que se rejoignent les intérêts des littéraires et des linguistes.

Le manuscrit et sa transcription

Suivant les auteurs, les pages de manuscrits sont plus ou moins complexes. Dans le cas du fonds Stendhal, la complexité des pages est très souvent poussée à son extrême, contenant des blocs de textes et des illustrations :

- dépendant de cadres rédactionnels différents (de la note diariste à l'essai théâtral ou littéraire),
- de mains différentes,
- à des dates différentes[4],
- dans des langues différentes[5],
- surchargés de biffes[6], ajouts, codes[7], surcharges[8]
- et enrichis de mise en forme graphique (gras, écriture en script, souligné…)

La page reproduite dans la fig. 1, page suivante, est représentative de cette complexité (plusieurs scripteurs, à plusieurs dates, différents types d'annotations et de corrections). Toute forme d'automatisation doit être écartée : les outils de reconnaissance optique de caractères ne sont bien évidemment pas en mesure d'analyser ce type de document ; les outils de délimitation d'unités textuelles restent encore des outils

2 Autographe : de la main de Stendhal.

3 Allographe : de la main d'autres personnes que Stendhal. Stendhal avait la fâcheuse habitude de dicter ses notes, y compris diaristes, à des proches ou des secrétaires, italophones pour certains. Il revenait ensuite sur ces prises de notes et les annotait.

4 Stendhal reprenait souvent des notes datant de plusieurs années pour les modifier ou les commenter

5 Stendhal pratiquait fréquemment le sabir, comme beaucoup de diplomates, intégrant dans son écrit des termes, expressions ou même des extraits complets en italien ou en anglais.

6 Biffes ou ratures

7 Codes : Stendhal utilise souvent des codes pour désigner des entités nommées : « Z » pour « Daru », son cousin, ou « 1000 ans » pour la ville italienne.

8 Surcharges : un premier mot est ébauché puis surchargé par un autre.

prototypiques et l'appropriation des interfaces logicielles de correction de ces délimitations n'est pas évidente pour un néophyte.

Par ailleurs, l'objectif n'est pas de fournir une transcription à l'identique qui respecterait exactement le positionnement, l'orientation et la taille des blocs de texte. A cette représentation diplomatique du manuscrit, nous avons préféré une représentation « pseudo-diplomatique » qui s'affranchit de certaines contraintes, trop coûteuses à décrire pour le transcripteur et peu compatibles avec un affichage en ligne.

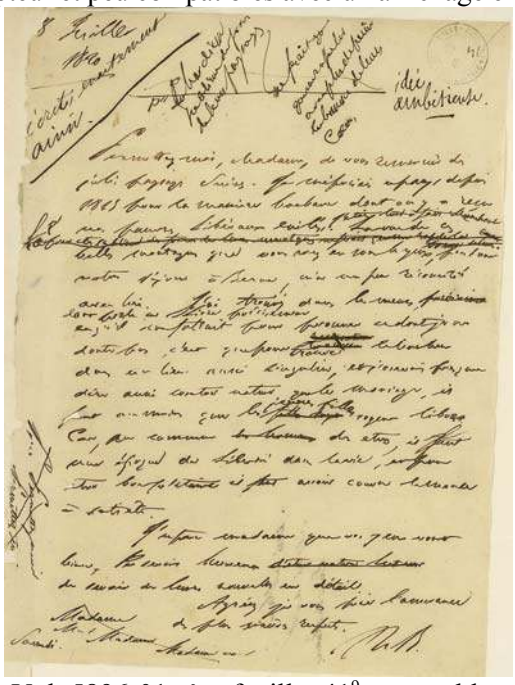

Figure 1 : Vol. 5896-01 rés., feuillet 41[9], rassemblant certaines des complexités de pages.

Afin de permettre toutes les observations et manipulation, la transcription doit donc être perçue comme une tâche à trois volets :

9 Image propriété de la Bibliothèque municipale de Grenoble

- identifier les blocs textuels et illustrations, c'est à dire les nommer dans la terminologie de spécialité,
- délimiter les blocs textuels identifiés,
- caractériser ces blocs en fonction de leurs propriétés.

Ces trois opérations indissociables que l'on observe dans de nombreux traitements de données langagières peuvent se faire dans un formalisme XML. La *Text Encoding Initiative* (TEI) propose à cet égard des solutions issues de nombreuses années d'expérience de linguistes et documentalistes. Toutefois, nous devions répondre à trois exigences majeures :

a) un délai de mise en place restreint par l'objectif de publication du premier tome des « Journaux et Papiers » fin 2010 ;
b) des compétences informatiques très variées des transcripteurs ;
c) une force de développement logiciel limitée ;

En conséquence, nous avons interprété la TEI comme un format d'échange (Burnard, 2001), c'est à dire non pas comme une norme d'encodage imposée mais comme un formalisme pivot vers lequel convertir un format plus adapté à la problématique.

L'objet transcrit est la page. Celle-ci se décrit d'une part en fonction de ses propriétés générales (identification, appartenance à un ensemble, propriétés codicologiques, identification du ou des transcripteurs et de la date de transcription, etc.), et d'autre part en fonction de son contenu.

Le contenu de la page est un ensemble de blocs textuels de différents types (pagination, foliotation, marginale, note de bas de page, ajout en marge) positionnés selon un découpage approximatif en 9 cases (combinaison des positions verticales haut-milieu-bas et horizontales gauche-centre-droite).

Chaque bloc textuel dispose d'un ensemble de propriétés, dont certaines dépendent du type de bloc, en particulier d'appartenance à des domaines (ex : *journaux et papier, correspondance, théâtre, etc.*), des corpus littéraires (ex : *« Vies de Haydn, Mozart et Métastase », « Le Rose et le Vert »*). Il est constitué de paragraphes, eux-mêmes constitués de lignes. Les éléments de correction et de mise en forme sont intrinsèques à la ligne.

Tous les éléments de structuration permettent trois niveaux de commentaires :

a) le commentaire scientifique, matériau préparatoire à l'édition critique format papier ;
b) le commentaire grand public, pour l'accompagnement du lecteur pour l'édition en ligne ;
c) le commentaire privé, à l'usage des transcripteurs, utilisant ainsi la transcription comme support de communication entre les membres de l'équipe afin d'élucider certains points complexes.

Le principe de délimitation d'unités organisées hiérarchiquement que permet XML s'avère cependant inadapté dans certains cas (Barnard et al., 1995). Ainsi, un paragraphe peut courir d'une page à une autre (voire un titre sur une double page), une biffe peut s'étaler du milieu d'une ligne au milieu d'une autre ligne. Des solutions telles LMNL (Caton, 2005) contournent élégamment ce problème en permettant le chevauchement de blocs délimités, mais ces langages sont généralement peu dotés en logiciels d'édition, ce qui rentre en conflit avec les contraintes de moyens de développement logiciel et de compétences informatiques des transcripteurs. Le document multistructuré (Portier *et al*., 2009) quant à lui permet de superposer différentes structures hiérarchiques sur un même document mais là encore, l'appropriation conceptuelle n'est pas à la portée de tous nos utilisateurs. Nous avons par conséquent opté pour des solutions simples : la biffe sur plusieurs lignes est une suite de lignes, chacune partiellement ou intégralement biffée ; les éléments découpés (paragraphe ou titre sur deux pages, ajout en interligne qui se prolonge en ajout en marge…) sont regroupés par le truchement d'un attribut de type pointeur.

La grammaire de description (la DTD) s'est construite grâce à un dialogue interdisciplinaire intensif fondé sur deux aspects fondamentaux :

- les explications bidirectionnelles : chaque discipline devant comprendre les besoins et contraintes de l'autre, mais aussi s'approprier la terminologie de l'autre ;
- les concessions bidirectionnelles : chaque discipline acceptant de s'alourdir un peu la tâche plutôt que d'alourdir exagérément la tâche de l'autre.

Le principe que nous nous étions fixé et que nous avons préservé est celui d'une « description sémantique » à la porté des utilisateurs. Les termes de description de la transcription - les balises et attributs de la DTD – sont ceux du domaine, compréhensibles et sans ambigüité pour le transcripteur, mais aussi en conséquence interprétables par le linguiste à destination duquel la transcription n'est pas initialement prévue.

Intégrée dans un logiciel libre d'édition de textes structurés et accompagnée d'une feuille de style permettant un affichage proche du résultat attendu, cette grammaire permet à une vingtaine de transcripteurs de travailler dans des conditions convenables : guidés par l'outil sur ce qu'ils peuvent faire et alertés lors d'erreurs ou de données absentes, ils travaillent avec un éditeur « à la Word » tout en produisant un document enrichi interprétable par le spécialiste comme par la machine.

Les 2.000 premières pages transcrites sont en cours de validation scientifique[10] et seront à disposition du public lors de l'inauguration[11] du site www.manuscrits-de-stendhal.org qui offre notamment des outils de recherche et d'extraction de pages par critères. Les résultats sont affichés en mettant les numérisations en vis-à-vis de leurs transcriptions pseudo-diplomatique (à l'image de la page) ou linéarisée (la résultante du processus d'écriture).

Dématérialisation et restructuration

L'objet de la transcription est la page. Toutefois les pratiques scripturales de Stendhal et le classement en volumes effectué par les premiers conservateurs génèrent un désordre inextricable que seuls quelques spécialistes peuvent démêler par un travail de longue haleine.

10 Un protocole de validation scientifique des transcriptions par les pairs (relecture par un collègue, corrections, validation par un comité) a été mis en place afin d'assurer une qualité scientifique optimale.

11 Les Manuscrits de Stendhal en Ligne seront inaugurés le 26 novembre 2009 à l'université Stendhal de Grenoble, ville natale de l'écrivain, sous l'égide des autorités culturelles, politiques et scientifiques locales et régionales.

Afin d'étudier les processus d'écriture chez Stendhal, d'un point de vue littéraire ou d'un point de vue linguistique, il est souvent impératif de se défaire de l'objet matériel « page de manuscrit ».

Qu'il s'agisse de répondre à « quelles sont figures, de la main de Stendhal, représentant des plans, classées selon les lieux décrits, puis par date de rédaction » ou à « quels sont les paragraphes comportant des biffes et des ajouts, classés par type d'écrit, puis pas ordre chronologique », la résultante est l'extraction à partir de l'ensemble des transcriptions, d'éléments vérifiant un certain nombre de propriétés et ordonnés selon d'autres propriétés.

L'organisation hiérarchique des transcriptions (un ensemble de pages, constituées de blocs de texte, eux-mêmes constitués de paragraphes, …) et le principe d'héritage implicite des propriétés dans l'arborescence XML, permettent de mettre en place ce procédé d'extraction.

L'intérêt réside non pas dans la faisabilité du processus mais dans son faible coût calculatoire, ce qui permet de l'intégrer sur une plateforme en ligne, donnant la possibilité à tout utilisateur de restructurer dynamiquement l'ensemble documentaire à sa guise.

La difficulté, en revanche, réside dans la conception d'interfaces logicielles permettant ce type de requêtes. Imposer un langage formel de requêtes, même simplifié, rebuterait une partie non négligeable des utilisateurs potentiels. Un système de requêtes en langue naturelle serait coûteux en développement et ne garantirait pas des taux de précision ni de décision suffisamment fiables.

L'histoire, pourtant courte, des technologies de l'information, montre que les interfaces novatrices pour une même fonction perturbent l'utilisateur (Ihadjane *et al.*, 2008). L'utilisateur n'effectue plus une recherche sur Internet, il « googlise ». Cette expression retient non seulement le fait qu'il va énumérer des termes et non formuler une question, mais aussi qu'il procédera à un tri parmi les réponses données pour y trouver le document qu'il recherche, que ce soit sur Google ou sur un autre moteur de recherche.

Nous partons du principe qu'il en est de même pour une interface d'extraction complexe : l'utilisateur ne doit pas être perturbé par une interface trop innovante. Par ailleurs, la pratique d'extraire des objets en fonction de leurs propriétés ou de propriétés de leurs constituants est monnaie courante dans des applications grand public : filtres de classement automatique des courriels, listes intelligentes de lecteurs multimédia.

La restructuration dynamique n'est un processus pertinent et fonctionnel que grâce à la qualité du travail de transcription (et *a fortiori* de la validation scientifique des transcriptions). Le principe de requêtes s'abstrayant de l'objet et l'objet lui-même étant extrêmement riche, la combinatoire des recherches, et par conséquent celle des résultats, est particulièrement vaste.

Notons toutefois que la dématérialisation de la page afin de recréer des ensembles textuels cohérents est une forme d'approche du manuscrit. Il ne faut pas exclure l'étude de la page en elle-même. L'outil informatique, fondé sur la description sémantique des constituants de la page, n'est là que pour faciliter l'accès à d'autres représentations, d'autres structurations de l'ensemble documentaire.

Il reste néanmoins la difficile tâche de qualifier ces résultats de restructurations dynamiques. Dans quelle mesure peuvent-ils être qualifiés de corpus ou de sous-corpus ?

Lectures rhizomatiques et corpus littéraire

La notion de corpus littéraire est celle donnée par la première acception du terme dans le Trésor de la Langue Française[12] :

> A.− PHILOL., SC. HUM. Recueil réunissant ou se proposant de réunir, en vue de leur étude scientifique, la totalité des documents disponibles d'un genre donné, par exemple épigraphiques, littéraires, etc.

Dans le cas des manuscrits de Stendhal, le corpus n'est pas défini *a priori*, mais se construit de manière empirique : il s'agit

12 http://www.cnrtl.fr/definition/corpus

d'un ensemble de documents qui ont un lien, en général mais pas exclusivement thématique.

C'est l'observation et l'analyse circonstanciée des manuscrits qui permettent de rassembler un ensemble de documents sous un intitulé commun[13] qui permet de désigner le corpus. La récurrence d'un aspect thématique fait émerger la notion de corpus auquel sont adjoints les documents au fur et à mesure de l'analyse. Ce regroupement permet notamment d'aller au delà de regroupements matériels ou typologiques : le corpus « Philosophia Nova » correspond aux différents documents du projet de livre de Stendhal ; les supports papiers (cahier cousus, feuillets libres…) et les natures de documents (notes de lecture[14] avec ou sans commentaires, notes personnelles, brouillons de textes) diffèrent mais l'objet thématique est le même : le matériau préparatoire à l'écriture d'une œuvre en particulier.

Les manuscrits de Stendhal sont à ce jour catégorisés selon 70 corpus différents : « Textes sur l'art », « Journaux », « Letellier »… Ce nombre augmente au gré des analyses des transcripteurs, mais tend à ralentir sa progression. Cette catégorisation est parfois multiple et hiérarchisée, correspondant à une granularité de précision et d'affinement des regroupements de textes. Ainsi, le corpus « Journal de 1808 » est un sous-corpus de « Journal de sa Vie depuis le 15 Avril 1806 Jusqu'au 3 Mai 1810 »[15], lui-même sous-corpus de « Journal ».

Le corpus, au sens littéraire, est par conséquent un mode de catégorisation et de caractérisation d'ensembles d'éléments textuels se rapportant à un même thème et à un type d'écrit. Il devient par conséquent une clé d'accès aux manuscrits pour les utilisateurs. Mis en opposition aux volumes

13 NDA : il s'agit ici de l'observation faite par le « naïf » en matière de recherche littéraire, « naïf » immergé dans les débats de l'équipe littéraire. Il ne s'agit en aucun cas d'une critique méthodologique.

14 La note de lecture consiste le plus souvent en un recopiage d'une œuvre existante.

15 Dans ce cas particulier, le nom de corpus est un intitulé donné par Stendhal lui-même.

et registres physiques conservés à la bibliothèque, ils offrent des accès thématiques compréhensibles pour le grand public comme pour les spécialistes.

Toutefois, nous constatons que l'utilisation du corpus comme point d'accès, comme méthode de regroupement des textes a fait dévier le sens initial du terme : de nouveaux corpus ont été définis qui ne sont plus thématiques et ne correspondent plus à proprement parler à des ensembles de documents. Ainsi le corpus « notes de relecture » désigne non pas un axe thématique mais la nature de l'objet considéré, l'objectif étant de regrouper virtuellement toutes les notes de lecture pour les comparer, voir quels textes Stendhal relit et annote de préférence. Ce glissement de l'usage premier défait alors le principe hiérarchique présenté précédemment et transforme la liste des corpus auxquels se rattache un objet textuel en une structure de traits, le même objet textuel peut alors se retrouver dans plusieurs corpus distincts recomposés virtuellement.

Le fonds peut être considéré comme un texte rhizomatique (*rhizome text*) au sens donné par Miles (1995) :

> a non-linear text that may or may not have various central 'nodes' joined in multiple ways with other text spaces, these links are understood to be thematically determined and defined.

La restructuration dynamique, autorisée par la dématérialisation numérique, offre alors un accès instantané et à l'envi à des lectures rhizomatiques des manuscrits. Le lecteur accède à un ensemble d'éléments textuels linéarisés selon un thème alors qu'ils sont matériellement éparpillés et sans hiérarchisation.

De surcroît, l'annotation diachronique offre d'autres formes de lecture du texte, au travers de ses mutations et évolutions dans le temps. Cette représentation des strates temporelles d'écriture est à la fois support d'analyse génétique et de démonstration.

Enrichi d'annotations elles-mêmes hypertextuelles et éventuellement multimédia, le corpus devient support

pédagogique aussi bien pour l'enseignement secondaire[16] et universitaire que pour l'amateur curieux.

Un corpus (pour la) linguistique

Si nous nous en tenons à la définition la plus reconnue de Sinclair :

> A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language. » (Sinclair, 1996)

le fonds des transcriptions, enrichi et annoté, ne constitue pas un corpus au sens strict, car il n'est ni sélectionné ni ordonné selon des critères linguistiques.

En revanche, dans son ensemble ou par sous-ensemble sélectionnés et ordonnés, il peut être utilisé comme un échantillon, non pas de la langue, mais d'une langue que l'annotation critique et génétique permet de caractériser selon plusieurs traits (outre son contenu purement textuel):

- qui sont les scripteurs
- quand le contenu a été rédigé (et corrigé)
- de quel type d'écrit il s'agit

L'ensemble peut être qualifié « d'écrit en genèse ». Des sous-ensembles (sous-corpus ?) peuvent être sélectionnés et ordonnés selon ces traits, grâce aux outils de restructuration dynamique. Il est par conséquent possible d'étudier d'un point de vue linguistique les pratiques scripturales sous plusieurs angles d'analyse. De nombreuses pistes de recherches peuvent être envisagées :

Ratures et réécritures

Est-il possible de caractériser les réécritures apportées par Stendhal à ses premiers jets (ou ses premières dictées à son secrétaire) ? Peut-on définir une typologie des réécritures? Est-il possible d'effectuer un rapprochement, d'observer des similarités entre les corrections à l'écrit et les phénomènes de

16 Les notions de genèse de l'œuvre littéraire font dorénavant partie des programmes de français en lycée.

disfluences[17] à l'oral (Blanche-Benveniste *et al*., 1990) ? Ces corrections sont-elles d'ordre différent suivant le type d'écrit (diariste ou littéraire) ? Les types de réécritures évoluent-ils dans le temps ?

Ces phénomènes sont généralement étudiés dans le cadre de l'apprentissage de l'écrit (Plane, 2003 ; Calil 2003), plus rarement dans le cadre de manuscrits d'écrivains[18] (Grésillon et Lebrave, 1982 ; Lebrave, 1983). Quand les ratures et réécritures sont étudiées transversalement à un groupe, la masse de données d'observations est générée par le nombre de membre du groupe. Quand elles sont étudiées sur un seul scripteur, il s'agit plus souvent d'analyses sur de petits volumes de données ce qui limite le potentiel généralisateur de la linguistique descriptive.

Dans les deux cas, les observations et analyses sont faites de manière synchronique, le ou les scripteurs sont étudiés à un instant *t* de leur vie. Pourtant, le processus de création écrite (voire littéraire) est généralement perçu sous l'angle de la cognition[19], capacité humaine fondamentalement évolutive. Les manuscrits de Stendhal tels qu'ils sont transcrits sont peut-être les premiers à offrir le matériau diachronique électronique nécessaire à une étude de la variation dans le temps d'un processus linguistique récurrent.

17 Claire Blanche-Benveniste définit la disfluence comme un énoncé où « le déroulement syntagmatique est brisé ».

18 Le manuscrit n'est certes pas l'apanage de l'écrivain, la rature non plus, mais rares sont les manuscrits et « brouillons » de personnalités de moindre rang qui soient préservés pour la postérité. A noter toutefois les « Carnets de Canuts » conservés par la Bibliothèque municipale de Lyon – les Canuts n'étant pas reconnus individuellement mais en tant que groupe à l'origine des premiers grands mouvements sociaux du XIXe siècle.

19 L'Institut des Textes et Manuscrits Modernes (ITEM), laboratoire CNRS souvent présenté comme précurseur dans le domaine de l'analyse linguistique des manuscrits affiche clairement cette perspective par une équipe « Manuscrit – Linguistique – Cognition ».

Sabir

Le sabir, en tant « langue mixte […] née du contact de communautés linguistiques différentes »[20], est une pratique courante chez Stendhal. Le contact des langues est dû aux fonctions diplomatiques de l'auteur. Mais ce sabir apparaît diversement dans les différents corpus (au sens littéraire du terme). Ces phénomènes ont été largement étudiés en tant qu'intérêt de Stendhal pour les Langues (Corredor, 2007).

D'un point de vue linguistique, c'est plus généralement le phénomène général de contact des langues qui est largement étudié. Mais il s'agit là de pratiques langagières relevant de l'oral. Le pluricodisme (le fait de combiner plusieurs « codes »), et le code-switching (alternance de code linguistique) sont généralement observés et caractérisés dans les pratiques orales. Ces pratiques langagières orales sont l'apanage des communautés baignant dans un environnement multilingue. Il est pourtant notoire que les polyglottes ont recours à ces mêmes pratiques à l'écrit, soit dans le cadre de communications épistolaires, soit dans le cadre de notes personnelles.

Les manuscrits de Stendhal offrent un cadre inédit d'observation de ces pratiques scripturales. Un parallèle doit être fait avec les mêmes pratiques à l'oral afin d'observer les similarités de pratiques mais aussi de tenter de circonscrire d'éventuelles divergences.

Dans cette approche, il est toutefois nécessaire de prendre en compte le personnage de Stendhal et l'environnement socio-culturel dans lequel il évolue, mais aussi les types d'écrits concernés par ce sabir. Le Stendhal qui étudie le « Character of Myself » (Journal du 20 janvier 1812, édité dans les Œuvres Intimes p. 819) n'est pas le même qui use du sabir à outrance dans son œuvre inédite « Earline ». L'observation et l'analyse des pratiques langagières d'un scripteur sont indissociables des raisons pour lesquelles et des conditions dans lesquelles il écrit : s'agit-il de codage de l'information par un diplomate, de pédanterie d'un intellectuel friand des jeux de langue ou d'une pratique courante dans certains cercles de la haute société du début du XIX^{e} siècle ? Et

20 Défintion du TLFI http://www.cnrtl.fr/lexicographie/sabir

surtout, d'un point de vue purement linguistique, quels sont les mécanismes syntaxiques, lexicaux et énonciatifs mis en œuvre et sont-ils similaires à ceux observés à l'oral ?

Orthographe désuète et erreurs d'orthographe

La rigueur orthographique n'était pas l'apanage de Stendhal. Dans la lourde tâche de transcription, les spécialistes stendhaliens s'efforcent de respecter l'écrit original. Toutefois, les éditions complémentaires numérique et papier doivent respecter les normes orthographiques modernes, ne serait-ce que pour les systèmes de recherche par mots clés. Les transcriptions portent ainsi la trace des variations entre l'orthographe originale et l'orthographe contemporaine.

Ces variations par rapport à la norme contemporaine sont de plusieurs ordres : orthographe désuète (tels les adverbes se terminant en « -ens »), influence néfaste des langues avec lesquelles le scripteur est en contact (« *filosofie »), erreur orthographique ou erreur grammaticale. Cette typologie sommaire reste à affiner. Sa compréhension précise présente un double intérêt : d'une part, permettre une édition qui offre une orthographe modernisée tout en respectant les erreurs orthographiques et grammaticales du scripteur ; d'autre part identifier et caractériser les influences du contact des langues – éléments complémentaire des phénomènes de sabir présentés dans la page précédente.

Ces pistes ne sont pas exhaustives et surtout ne peuvent pas être toutes creusées simultanément par une seule équipe. Pour ces raisons, l'ensemble des transcriptions, à l'exception des annotations critiques prévues pour l'édition papier, est mis librement[21] à la disposition de la communauté. Nous ne sommes pas face à un corpus de « plusieurs millions de mots » (Sinclair, 1996) – nous ne sommes pas en mesure que quantifier le corpus à ce jour, si ce n'est en nombre approximatif de pages, les transcriptions étant en cours.

21 Licence CreativeCommons BY-NC-SA (reconnaissance de paternité de l'œuvre créée, diffusion non commerciale et dans des conditions identiques).

Ces analyses, par ailleurs, ne révélerons de propriétés générales de la langue mais uniquement les particularités d'un auteur du XIX^{e}.

Les « Manuscrits de Stendhal », constitués en tant que corpus numérique, apportent toutefois un autre regard méthodologique sur ce type d'analyses. Mis librement à la disposition du public, le corpus n'est plus confiné dans un laboratoire, dans une école de pensée linguistique. Annotés et enrichis, les écrits sont contextualisés pour le linguiste non-stendhalien et lui apportent les outils d'interprétation de ses observations. Accompagnés d'outils d'extraction et de réorganisation (la sélection et l'ordonnancement de la définition Sinclairienne de la linguistique de corpus), le corpus permet des analyses quantitatives et qualitatives de phénomènes linguistiques indépendamment des intentions initiales de ses concepteurs. Enfin, décrit finement et mis en vis-à-vis des images des pages manuscrites numérisées, il offre un accès numérique (de sélection et d'ordonnancement) à la langue et à la métalangue de manière lexicale (par la description) et visuelle (par l'image).

Discussion conclusive

La démarche originale d'une collaboration entre linguistes et littéraires pour la constitution d'un corpus littéraire et linguistique est en soi une expérience scientifique pluridisciplinaire digne d'intérêt mais transige avec les fondements théoriques et méthodologiques de la linguistique de corpus telle qu'elle a pu être définie et mise en pratique par l'école sinclairienne.

Constituer un corpus dans un dialogue interdisciplinaire ne peut s'avérer constructif que dans la recherche d'intérêts réciproques : apporter mais aussi et surtout expliquer les méthodes et outils de la linguistique de corpus à une communauté scientifique qui en présente le besoin ; en contrepartie, bénéficier de la constitution d'un corpus annoté et enrichi dont la fiabilité et la qualité sont garanties par un processus de relecture et de validation scientifique et technique du corpus et de ses informations paratextuelles.

Le regard et l'analyse du linguiste sont une plus-value non négligeable à la compréhension du personnage stendhalien. Sans pour autant viser l'étude psychanalytique – pratique (voire déviance) courante dans les travaux sur la relation de Stendhal à la langue[22] – la linguistique appuyée par une démarche en corpus offre un arsenal d'outils d'interprétation et de justification des pratiques scripturales stendhaliennes. Inversement, le regard et l'analyse du littéraire apportent les informations nécessaires de contextualisation des analyses linguistiques. Ainsi, interpréter le sabir stendhalien et sa variation ne peuvent se faire qu'en connaissance des lieux et contextes de rédaction.

En revanche, si l'acception littéraire du terme corpus ne peut être remise en question et offre par ailleurs au linguiste une typologie des écrits transcrits, l'acception linguistique du terme est fortement malmenée. Dans la démarche du linguiste de corpus, il est d'usage de définir le phénomène linguistique à étudier puis de constituer un corpus représentatif, ou d'utiliser un corpus de référence. Dans le cas de cette expérience, nous nous trouvons dans la situation incongrue où le corpus n'est pas de référence, pas représentatif d'une communauté d'usage, mais préexiste tout projet d'étude linguistique, tout en étant riche de phénomènes langagiers identifiés et délimités.

Le terme de corpus, dans le sens de Sinclair, n'est donc pas applicable *stricto sensu* à l'ensemble textuel du fonds Stendhal. Cependant, la structuration et l'enrichissement apportés par les besoins de description littéraire, cadrés par un formalisme rigoureux, ont permis de construire un matériau langagier conséquent représentatif des pratiques scripturales de Stendhal dont les phénomènes analysables peuvent être extraits automatiquement.

Nous ne sommes pas dans le cas de la construction d'un corpus comme échantillon représentatif de la langue, mais dans le cas d'un corpus existant préalablement aux études linguistiques. Sa caractérisation circonscrit les observations qui peuvent être faites.

22 Voir par exemple le chapitre « Parole du moi, parole des autres » de (Crouzet, 1981), pp. 21-80.

Le dialogue interdisciplinaire est la pierre angulaire de cette construction. Fondé sur la mise en commun des besoins et attentes, qui se sont avérées complémentaires, il a permis la mise en place de méthodologies et outils informatiques communs, les travaux des uns formant la base de travail des autres.

S'affranchissant ainsi de la matérialité de la page de manuscrit, le chercheur littéraire dispose d'un outil permettant l'observation de la genèse d'une œuvre, le linguiste d'un microscope permettant l'analyse des pratiques scripturales.

Références bibliographiques

Barnard D., Burnard L., Gaspart J.-P., Price L., Sperberg-McQueen C.M., Varile G. (1995). « Hierarchical Encoding of Text : Technical Problems and SGML Solutions », *Computers and the Humanities*, 29, Kluwer Academic Publishers.

Blanche-Benveniste C., Bilger M., Rouget C., Van Den Eynde K. (1990). *Le français parlé : Etudes grammaticales.* Paris : CNRS Editions, « Sciences du langage ».

Burnard L. (2001). XML+TEI, un mariage fait aux cieux, présentation à la MSH de Lyon, novembre 2001, [en ligne], http://www.tei-c.org/Talks/lyon-3.ppt, consulté le 23 avril 2009.

Calil E. (2003). « Processus de création et ratures : analyse d'un processus d'écriture dans un texte rédigé par deux écolières », *Langage et Société* 103 : 31-51.

Caton P. (2005). « LMNL Matters », *Extreme Markup Languages*. Montréal, Quebec.

Corredor M.-R. (2007). *Stendhal à Cosmopolis : Stendhal et ses langues*, [textes réunis et présentés par]. Grenoble : Ellug.

Crouzet M. (1981). *Stendhal et le langage*. Paris : Gallimard, NRF.

Grésillon A., Lebrave J.-L. (1982). « Les manuscrits comme lieu de conflit discursifs », *in La genèse des textes, les modèles linguistiques*. Paris : Editions CNRS.

Ihadjane M., Chaudiron S. (2008). « Quelles analyses de l'usage des moteurs de recherche ?», *in* B. Simmonot (éd.), *Questions de Communication* 14.

Lebrave J.-L. (1983). « Lecture et analyse des brouillons », *in* A. Grésillon et J.-L. Lebrave (eds), *Langages* 69 : 11-23.

Meynard C. (2007). *Stendhal, Histoire d'Espagne*, inédit présenté et annoté par Cécile Meynard. Paris : Kimé, coll. « La Chasse au Snark ».

Miles A. (1995). « The Hypertext project : Our working understanding of Hypertext », *in* RMIT Communication Studies, [en ligne] consulté le 25 avril 2009, http://hypertext.rmit.edu.au/publications/one/hypertext_terms.html

Plane S. (2003). « Stratégies de réécriture et gestion des contraintes d'écriture par des élèves de l'école élémentaire : ce que nous apprennent des écrits d'enfants sur l'écriture », *Rivista Italiana de Psicolinguistica Applicata*, anno III/1 : 57-77.

Portier P.-E., Calabretto S. (2009). « Modélisation des connaissances dans le cadre de bibliothèques numériques spécialisées », *in* J.-G. Ganascia et P. Gançarski (eds), *Extraction et Gestion des Connaissances (EGC)*, Revue des Nouvelles Technologies de l'Information RNTI. Strasbourg : Cépaduès-Editions pp. 391-396.

Sinclair J. (1996). « Preliminary Recommendations on Corpus Typology », EAGLES documents, [en ligne], consulté le 25 avril 2009
http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html

Spengler H. (2006). « Imaginaire et écriture de l'énergie dans l'œuvre de Stendhal : "La Révolution entre dans la littérature" », thèse de doctorat de l'Université Grenoble 3.