CLELIA :
Développement d'un corpus littéraire et linguistique par le biais d'un dialogue interdisciplinaire

Thomas Lebarbé
Laboratoire LIDILEM, EA 609, Université Stendhal – Grenoble 3

La Bibliothèque de Grenoble possède la quasi totalité des manuscrits de Stendhal. Cet ensemble de plus de seize mille feuillets compte parmi les plus importants des fonds de manuscrits littéraires modernes et par sa taille et par la présence de textes majeurs, entre autres *Lucien Leuwen,* ou *Vie de Henry Brulard.*

Cependant, à côté de ces manuscrits figurent des séries aléatoires de correspondances, de brouillons et ébauches de pages de journal, de pensées, de papiers divers, etc. Aujourd'hui, le classement du fonds n'est donc satisfaisant ni du point de vue chronologique ni du point de vue thématique.

Il est par ailleurs essentiel de préserver ce patrimoine précieux mais aussi de le valoriser en le mettant à disposition du grand public aussi bien que des chercheurs. Pour ce faire, depuis 1996, la Bibliothèque municipale s'est lancée dans une vaste campagne de numérisation et de mise en ligne des manuscrits.

Dans le même temps, une équipe de chercheurs de l'Université Stendhal-Grenoble 3, dirigée par Gérald Rannaud, a conçu un prototype de base de données associant les images des manuscrits, leurs transcriptions et des informations sur les pages.

Depuis l'an dernier, le projet est entré dans une nouvelle phase, sous le nom de CLELIA – Corpus Littéraire et Linguistique assisté par Intelligence Artificielle, grâce à la collaboration entre une équipe littéraire (dirigée par Cécile Meynard) et une équipe de recherche en informatique – linguistique (dirigée par Thomas Lebarbé). Dans cette communication, nous montrerons la plus-value d'une telle approche interdisciplinaire.

L'objectif était de créer une application en ligne qui permette différents niveaux d'accès (du grand public au chercheur spécialisé) aux images de manuscrits, à leurs transcriptions et aux informations afférentes. Le but est également de faciliter le travail des transcripteurs qui déposent leurs transcriptions en ligne (et les visualisent donc plus facilement) et les font valider par un comité scientifique ; et de préparer des éditions papier.

Nous avons d'abord défini ensemble un modèle de structure documentaire beaucoup plus évolué que la simple table de la base de donnée initiale. Après un an de travail et une vingtaine de versions du modèle documentaire progressivement affiné, nous avons abouti à :

- un modèle de document en XML permettant une description détaillée des pages
- une base de donnée appariée à la structure XML afin d'optimiser les outils de recherche d'information
- un logiciel libre utilisé pour les transcriptions, donnant un aperçu en temps réel des fiches saisies

- une plateforme en ligne permettant le dépôt, la validation et la consultation des transcriptions et conçue de manière modulaire pour greffer des outils avancés : enrichissement morphosyntaxique des transcriptions, visualisation du corpus sous forme de cartographies…

Cette plateforme est un magnifique outil de valorisation des Manuscrits de Stendhal, dont le principe est toutefois adaptable à tout autre corpus et/ou auteur.

Grâce au dynamisme des deux équipes, le projet est bien avancé et la plateforme sera en ligne courant 2008. Ce travail interdisciplinaire a été efficace grâce à deux conditions nécessaire : 1) la régularité et la bidirectionnalité du dialogue interdisciplinaire ainsi que l'intercompréhension qui en résulte ; 2) le principe de réciprocité, chaque équipe bénéficiant des outils et ressources produits par l'autre équipe.

Du point de vue linguistique, grâce au travail de transcription des Stendhaliens, nous disposerons rapidement d'un corpus relativement volumineux aux propriétés suivantes :

- chaque unité textuelle est datée, donnant ainsi un corpus diachronique
- le processus de rédaction y est représenté de manière structuré et chronologique
- les erreurs (orthographe, grammaire, etc.) des différents scripteurs (Stendhal dictait souvent ses pensées à des secrétaires et faisait relire ses écrits par ses amis).

Le corpus littéraire des manuscrits de Stendhal représente donc aussi un corpus linguistique support de plusieurs études à la fois utiles pour la connaissance de la langue et pour la connaissance de l'œuvre de Stendhal :

- L'analyse syntaxique automatique de ce corpus, nécessaire à une meilleure indexation, sera un tour de force car elle devra prendre en compte les différentes incohérences orthographiques et grammaticales.
- L'étude des évolutions lexicales et stylistiques sur la période 1802-1842 donnera une meilleure connaissance des évolutions langagières de l'auteur. Par ailleurs, une collaboration de dessinant avec les équipes de recherche de Rouen, une étude similaire et contrastive sera menée sur les manuscrits de Flaubert.
- Enfin, Stendhal étant connu pour ses plagias de ses contemporains, nous souhaitons lancer un projet afin d'identifier automatiquement les éléments d'écrit concernés et les associant aux textes originaux, apportant ainsi aux chercheurs Stendhalien un support argumentatif pour le commentaire génétique et critique.