Intervention d’Y. Mosset et de T. Rainsford

La littérature médiévale avant le XIIIe siècle n'existe que sous la forme versifiée : la langue est informée par le vers. Au plan littéraire, quatre notions s'entrecroisent : les genres littéraires, les métriques, les formes syntaxiques et les auteurs. Deux auteurs de la fin du XII sont abordés : Béroul et Chrétien de Troyes.

Le projet *Syntactic Reference Corpus of Medieval French* repose sur un ANR. L'objectif était de produire un *treebank* (un corpus arboré), en ajoutant une couche d'annotation. Deux bases sont utilisées comme la Base de Français Médiéval à l'ENS de Lyon ou le Nouveau Corpus d'Amsterdam de l'Université de Stuttgart. L'annotation est à la fois morphosyntaxique, syntaxique, du discours direct et de la forme métrique (ici des débuts ainsi que des fins de vers).
Comment créer un corpus syntaxique? Il n'a pas toujours de modèle, notamment dans le français médiéval. Deux experts annotent le même texte via NotaBene. Deux versions étiquetées peuvent alors être comparées et discutées. L'exploitation est fondée sur Tiger-Search pour les requêtes. L'export pose problème sous un format peu connu Tiger-XML. Des scripts notamment groovy favorisent des résultats. L'interrogation reste difficile. Le logiciel demande un investissement en temps, au moins une journée. En moyenne, 1500 mots par jour et par personne peuvent être annotés.
Cinq textes sont utilisés pour questionner la limite de la phrase et la limite des vers. Des statistiques peuvent alors être menées avec une volonté contrastive entre les textes. Dans le temps, les métriques et les structures sont évolutives.
Le rejet est étudié à la fois de façon quantitative et qualitative. Les requêtes posent la question de cas identifiés : en effet, le chercheur enlève certains exemples. Les statistiques sont alors modifiées. Le logiciel ne permet pas d'identifier certains types de rejet, ce qui sous-entend la mise en place de dépouillement manuel partiel.
Le corpus annoté fournit un appui quantitatif sur une très large base de données à des constats qualitatifs et intuitifs. L'interrogation sous la forme de bases demande de préciser les définitions. Le mélange entre quantitatif et qualitatif a fonctionné sur un rebond. Si la définition est qualitative, cela permet de mieux interroger le corpus. Les statistiques sont un point de départ, mais il faut aussi dépouiller ensuite de façon qualitative. Le dialogue semble essentiel pour aboutir à un discours convaincant. Le temps reste la contrainte principale. Les cas éliminés à la main présentent des critères clairs : toute définition concrète peut être réinsérée sur le logiciel. Le tri manuel devient plus facile quand le nombre de cas se restreint.