Intervention de M. Decorde et Y.-F. Le Lay

Les SHS connaissent un mouvement d'internationalisation. Le texte reste une source de prédilection. Le texte est mis en valeur au sein d'une recherche communautaire et accumulative. Le nombre semble garant d'une objectivité, mais le chercheur a besoin d'un mélange entre quantitatif et qualitatif.

M. Decorde présente la plateforme TXM. TXM a été initié par une ANR textométrie avec le rôle de S. Heiden et de B. Pincemin. TXM est un outil *open source* d'analyse de données textuelles. L'hébergement se fait sur sourceforge.
Matthieu Decorde présente TXM avec ces deux corpus : la « Quête du Graal » et « Discours », un échantillon du corpus de Damon Mayaffre. Un corpus sur les reconquêtes des Berges à Lyon d'Emeline Comby va être utilisé pour faire la démonstration. Les corpus chargés apparaissent dans une colonne à gauche. La console présente les messages de la plateforme. Le corpus peut être visualisé via la fonction « information » : le moteur de recherche trouve toutes les informations sur le corpus. Sur chaque mot sont rajoutés le lemme (entrée du dictionnaire) et la morphosyntaxe (la fonction) via Tree Tagger. Le corpus est constitué de textes. Un texte par défaut a un nom. Des métadonnées peuvent être rajoutées pour faciliter les comparaisons. Le corpus peut en effet être importé ou être chargé, il lit plusieurs formats dont Alceste et Hyperbase. La textométrie est fondée sur le retour au texte via des éditions textuelles, donnant les métadonnées du texte et la lecture classique du texte. Le lexique permet de voir les mots : les fréquences correspondent au nombre de fois qu'ils apparaissent dans le corpus. Du lexique, il est possible d'aller à la concordance centrée autour d'un pivot et autour de références. L'édition numérique donne le lemme et la fonction. Les co-occurrences comptent les mots autour d'un mot choisi : le score. La démonstration porte sur l'exploration mais sont aussi présentes des requêtes de type CQL. La progression permet de classer les textes dans un ordre : la comparaison porte sur Rhône et Saône, pour montrer les évolutions dans le temps du mot Saône.
Après ces quelques fonctions, la parole est laissée à Yves-François Le Lay qui présente le lien entre TXM et R. Le corpus est numérisé en archives puis OCRisé. La structuration et l'import des données sont faits sous la forme .txt et .csv dans un même dossier et fondés sur une logique d'identifiant (dit id). Le corpus peut être partitionné pour avoir une entrée contrastive (temps, auteur, espace). L'exemple présenté est fondé sur une partition par année : tous les mots sont visibles mais il est possible de choisir les mots les plus spécifiques ou des seuils de fréquence. Les mots les plus spécifiques sous la forme d'un index sont gardés dans une table lexicale qui croise les mots en ligne et les parties en colonne. Cette table devient le point de départ des analyses statistiques suivantes : R peut être utilisé sous TXM (consciemment ou non) ou en faisant un export avec R. Des *packages* R peuvent alors être mobilisés, ici ADE4. L'AFC permet de voir la structuration dans le temps (F1) et dans l'espace (F2) de la reconquête des Berges : la projection des mots aide à la compréhension. R propose des *clustering* hierarchiques (ici classification ascendante hiérarchique) : 8 classes sont créées et chacune d'elle peut être étudiée. Le *cluster* est fondé sur la distance du khi2 entre les mots : deux mots se ressemblent d'autant plus qu'ils se ressemblent en termes d'années. Les classes peuvent être projetées sur un plan factoriel.