Présentation d'un chantier en cours : le développement d'un environnement de prétraitement des corpus textuels

Le *prétraitement* des données est une étape décisive du processus de recherche pour l'analyse du discours outillée (ou textométrie), qui médiatise son activité interprétative par des prises de mesure et de vue sur la matérialité textuelle.

Lors de cette étape, le chercheur procède à différentes opérations, parmi lesquelles la *correction* des fautes orthographiques et la *normalisation* des différentes graphies cohabitant dans le corpus pour un même « mot ». Si ces opérations font fortement écho aux pratiques philologiques anciennes (Rastier, 2001 ; Viprey, 2005), elles n'en sont pas moins l'occasion de choix herméneutiques.

Ces opérations, qui ne peuvent donc être confiées à un automate, sont particulièrement laborieuses et deviennent chronophages lorsque le chercheur est confronté à des données textuelles fortement *bruitées*.

En effet, certains textes, de par leur genre et/ou leur contexte de production, comportent un nombre important de formes graphiques pour une même unité. C'est par exemple le cas de corpus RSN (Twitter, Facebook), où abondent notamment les fautes de frappe et d'orthographe, ainsi que les phénomènes d'économie linguistique.

Dans d'autres cas, l'extrême bruitage des données résulte de la procédure visant à désinscrire le texte de son support initial, à savoir l'océrisation.

Afin de faciliter le traitement nécessairement semi-automatique et de rendre le prétraitement de ces données compatible avec le temps de la recherche, nous avons initié le développement d'un environnement dédié à assister le chercheur dans le prétraitement des données textuelles.

Dans le cadre de notre intervention, nous présenterons les principes sous-jacents au fonctionnement de cet outil et le cahier des charges de ses fonctionnalités. Ensuite, adoptant le point de vue de l'utilisateur, nous procéderons à une démonstration des « parcours de correction » soutenus par cet environnement logiciel.