Bien choisir ses données d'apprentissage pour le TAL en contexte multi-hétérogène : l'exemple de l'ancien français

Isabelle Tellier*

*Laboratoire LaTTiCe UMR 8094, Paris

Les techniques d'apprentissage automatique supervisé sont maintenant largement exploitées dans la communauté TAL, que ce soit pour la classification de textes, l'étiquetage morphosysntaxique ou la construction d'un analyseur syntaxique. Elles font partie de la "boite à outils" de tout chercheur du domaine. Mais ces méthodes requièrent pour être efficaces de grandes quantités de données manuellement annotées qui ne sont pas toujours disponibles, surtout si on se confronte à des textes écrits dans des formes "non standard" comme les SMS ou les tweets. Pour correctement traiter de tels textes, le problème n'est plus de disposer d'un bonne technique d'apprentissage mais de disposer de bons exemples. La problématique de la recherche s'est ainsi largement déplacée des programmes aux données, et nécessite de nouvelles approches pour sélectionner des données "sur mesure"' en fonction de l'application visée. Pour illustrer mon propos, je m'appuierai sur des expériences menées sur des textes datant d'une autre époque mais où la langue connaissait aussi une forte variabilité : le Moyen-Age !