MultiTAL — Plateforme de documentation et d'expertise des outils et ressources pour le traitement automatique des langues orientales et des langues peu dotées

Porteur : ERTIM (Mathieu Valette : mvalette@inalco.fr)
Membres : J.-M. Daube, M.-A. Moreaux, D. Nouvel, S. Mkhitaryan, D. Sadoun, F. Stuck

Il s'agit d'un projet de plateforme experte en matière de traitement technologique des langues mettant à la disposition de la communauté des connaissances actualisées et une expertise critique relatives aux ressources linguistiques (corpus, lexiques, etc.) et aux outils de traitement automatique des langues (programmes, composants logiciels, etc.) disponibles, en particulier (i) lorsqu'elles relèvent du domaine INALCO ; (ii) lorsqu'elles sont peu dotées.

La description des langues et les typologies linguistiques existantes ne recoupent pas celles requises pour leur outillage. Par exemple, des langues non apparentées peuvent nécessiter des outils similaires ou certains outils communs du point de vue de leur traitement automatique : l'estonien (langue finno-ougrienne, fennique) peu outillée est une langue agglutinante et compositionnelle au même titre que l'allemand (indo-européenne, germanique) nécessitant un analyseur morphologique reconnaissant les affixes et les composants internes des mots ; le chinois (sino-tibétaine) et l'anglais (indo-européenne, germanique) sont toutes deux des langues isolantes qui requièrent un regroupement des mots graphiques en unités de sens complexes. Par ailleurs, de multiples problèmes bloquants sont souvent méconnus ou sous-estimés dans le traitement informatique des langues, (encodage, sens de l'écriture, ligature, segmentation, etc.).

Une connaissance approfondie et exhaustive des outils de TAL selon un référentiel commun, permettrait non seulement l'adaptation et la spécification d'outils pour le traitement des langues moins dotées (ex. l'adaptation d'un étiqueteur morphosyntaxique statistique existant à une nouvelle langue) mais aussi des innovations en matière de méthodes et d'algorithmique utilisable sur toutes les langues a priori (ex. les modèle statistiques de traitement du corpus).

L'enjeu de la plateforme est double. Il s'agit en premier lieu de mettre à disposition de la documentation à jour (signalement des ressources linguistiques et des outils logiciels (désormais R&O) de leur localisation et description) et, lorsque cela est juridiquement et techniquement possible, des échantillons ou les R&O eux-mêmes. Cette documentation unifiée constitue la contribution essentielle de la plateforme. Elle résulte d'un examen critique et d'une expertise technique des R&O signalés ou mis à disposition – et de leur actualisation régulière.