Stabilité interlangue des catégories morpho-syntaxiques

Le projet "Universal Dependancies"[1] rassemble des corpus arborés de données linguistiques de plus d'une centaine de langues, annotées avec le même jeu d'étiquettes morphosyntaxiques (donc strictement le même ensemble de parties du discours, ce qui conduit par exemple à avoir une catégorie ADP (adposition) qui subsume les prépositions et les post-positions) et avec le même jeu de relations de dépendance. Un des aspects phare de ce projet est l'élaboration de ce jeu d'étiquettes et de relations unique qui a vocation à couvrir l'ensemble des langues connues.

Le but du stage est de mener un certain nombre d'études sur ces données pour comparer finement d'une langue à l'autre la réalité linguistique recouverte par ces catégories universelles. Différentes méthodes relevant du TAL seront utilisées dans ce travail.

Par exemple, on pourra produire les plongements lexicaux distributionnels statiques (type Word2vec ou GloVe) des mots associés à la catégorie ADJ en anglais et en chinois, et mesurer la proximité et l'homogénéité de ces plongements.

De même, on pourra entraîner un parseur sur, par exemple, les arbres syntaxiques des données du français dans lesquelles les adjectifs epithètes sont systématiquement réanalisés comme des subordonnées relatives et mesurer à quel point ce parseur a de meilleures performances sur les données de langue où la distinction entre épithète et relative est bien moins marquée comme en coréen ou en mandarin.

Le stage aura lieu au laboratoire Lattice (ENS, 1, rue Maurice Arnoux, Montrouge), et sera encadré par Mathieu Dehouck (CR CNRS) et Pascal Amsili (PR, Sorbonne Nouvelle).

1. *https://universaldependencies.org*