Phraseotext : annotation syntaxique et mise en ligne d'un corpus latin (stylistique et phraséologie)

Louis Autin, Kamel Bouzidi, Olivier Kraif, Julie Sorba

Phraseotext : annotation syntaxique et mise en ligne d'un corpus latin (stylistique et phraséologie)

Louis Autin*, Kamel Bouzidi**, Olivier Kraif** et Julie Sorba**
*Litt&Arts, TRANSLATIO et **LIDILEM - Université Grenoble-Alpes

Présentation du projet

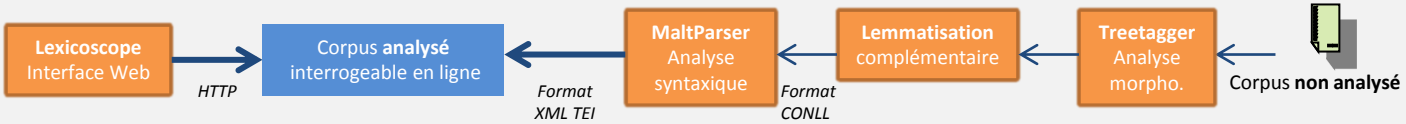
Projet AGIR-POLE

Objectif : Produire un corpus arboré de textes latins afin de les rendre accessible à travers une interface Web.

Historique : Le Lexicoscope, un outil issu du projet Emolex pour l'exploration de la combinatoire lexicale (Kraif & Diwersy, 2014).

Visée : Etudier la phraséologie dans une perspective textuelle.

Schéma de fonctionnement final



Constitution d'un corpus pour l'apprentissage d'un modèle (Malt) [état actuel du projet]

Corpus arborés préexistants : **Projet Perseus** LDT (Bamman & Crane, 2011), **Index Thomisticus** IT-Treebank (McGillivray et al., 2009), **PROIEL Treebank** (Haug et al., 2009).

→ **Harmonisation et fusion** (au format .conll) des différents jeux d'étiquettes (morphologie et syntaxe). Nettoyage des corpus.

→ Résultat fourni à **MaltParser** comme **corpus d'apprentissage** *(≈ 300 000 mots, 200 000 nécessaires pour obtenir un résultat pertinent)*.

↓

Evaluation des sorties du système **Treetagger/MaltParser** avec les différents jeux d'étiquettes établis plus tôt.

← **Etude des résultats, choix définitif** d'un jeu d'étiquettes, et **optimisation** des paramètres : moins de 10% d'erreurs sur le corpus test.

← **Intégration** au Lexicoscope.

Corpus final

Deux genres littéraires unis par leur caractère oratoire : les **rhéteurs** (Cicéron, Sénèque le Père) et les **historiens** (Salluste, César, Tite-Live, Tacite), pour un corpus de plus de 3 millions de mots.

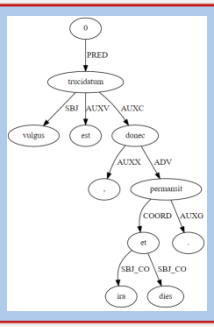
Auteur	Mots	Textes
César	98 611	2
Cicéron	1 594 938	30
Tacite	193 169	5
Tite-Live	789 059	13
Salluste	38 670	2
Sénèque	409 980	24
Sénèque le père	121 806	4
Total	3 246 233	80

Choix d'annotation

Reprise du schéma de **IT-Treebank** (conversion des étiquettes en majuscules, harmonisation de tags, par ex. SB → SBJ), simplification des relations composées de **Perseus LDT** (par ex. APOS_ExD0_PRED_CO ou NOM_ExD5_PRED_CO).

Exemple de sortie annotée :

« Vulgus trucidatum est, donec ira et dies permansit » (Tacite, *Ann.*, I, 68).



Points problématiques

Difficultés actuelles, inévitables dans un projet aussi neuf :

- avec **Treetagger** (morphologie) : structures elliptiques ; restitution désinentielle des abréviations ; cas des enclitiques (en cours de résolution) ;
- avec **Maltparser** (syntaxe) : héritage des mauvaises analyses morphologiques ; flottement dans l'analyse des structures complexes (par ex. : propositions infinitives ou participiales rarement étiquetées avec un sujet et un verbe).

Résultats escomptés

Recherche d'expressions complexes : Extraction des concordances d'une expression ou d'une construction, en posant des contraintes sur son environnement syntaxique.

Recherche de collocatifs fréquents : Extraction des lexicogrammes (tableaux de cooccurrences), contenant les collocatifs syntaxiques les plus significatifs.

Exemple de recherche : étudier la spécificité syntaxique et sémantique de l'infinitif de narration **tacitéen**, qui semble lié au collectif : le Lexicoscope permettra de relever tous les infinitifs de narration dont le sujet est la foule (*uulgus*, *multitudo*, *etc.*), mais aussi qui s'inscrivent dans un contexte où ces substantifs apparaissent.

Références : Kraif, O., & Diwersy, S. (2014). « Exploring combinatorial profiles using lexicograms on a parsed corpus: a case study in the lexical field of emotions ». Dans P. Blumenthal, I. Novakova, & D. Siepmann (éds.), *Actes du colloque international Nouvelles perspectives en sémantique Lexicale et en organisation du discours*. Osnabrück, Allemagne : Peter Lang // McGillivray, B ., Passarotti, M., Ruffolo, P. (2009). « The Index Thomisticus Treebank Project: Annotation, Parsing and Valency Lexicon ». TAL 50(2),103–127 // Bamman, D., Crane G. (2011). « The Ancient Greek and Latin Dependency Treebanks ». Dans C. Sporleder, A. van den Bosch and K. Zervanou (éds.), *Language Technology for Cultural Heritage*, ser. Foundations of Human Language Processing and Technology, Springer, 2011 // Haug, D.T.T., Jøhndal, M. L., Eckhoff, H. M., Hertzenberg, M. J., Müth, A. (2009). « Computational and Linguistic Issues in Designing a Syntactically Annotated Parallel Corpus of Indo-European Languages ». TAL 50(2), 17-45.