



**HAL**  
open science

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

Louis Autin, Kamel Bouzidi, Olivier Kraif, Julie Sorba

### ► To cite this version:

Louis Autin, Kamel Bouzidi, Olivier Kraif, Julie Sorba. Phraseotext : annotation syntaxique et mise en ligne d'un corpus latin (stylistique et phraséologie). Humanités numériques et Antiquité, Sep 2015, Grenoble, France. . hal-01844375

**HAL Id: hal-01844375**

**<https://hal.science/hal-01844375>**

Submitted on 19 Jul 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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



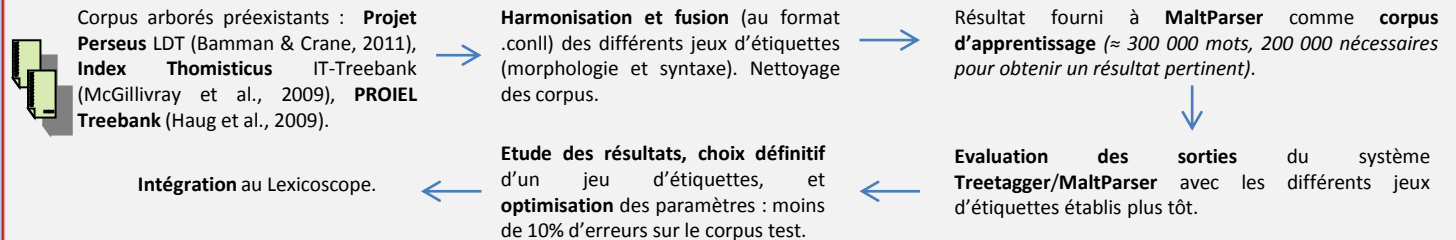
Louis AUTIN\*, Kamel BOUZIDI\*\*, Olivier KRAIF\*\* et Julie SORBA\*\*  
 \*Litt&Arts, TRANSLATIO et \*\*LIDILEM - Université Grenoble-Alpes

## SCHÉMA DE FONCTIONNEMENT FINAL



## CONSTITUTION D'UN CORPUS POUR L'APPRENTISSAGE D'UN MODÈLE (MALT)

[état actuel du projet]



### 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.

### 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.

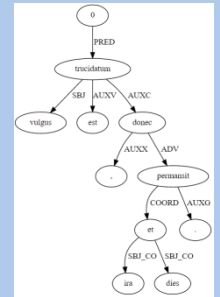
Auteur	Mots	Textes
César	98 611	2
Cicéron	1 594 938	30
Tacite	193 169	5
Tite-Live	789 059	13
Salluste	38 670	2
Sénèque	409 980	24
Sénèque le père	121 806	4
<b>Total</b>	<b>3 246 233</b>	<b>80</b>

### 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 **taciteén**, 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.