

## **Dominique LONGRÉE**

*Lemmatising and tagging Latin literary classical texts for more than 50 years: The LASLA experience...*

Since 1961, the Laboratoire d'Analyse Statistique des Langues anciennes ([www.ulg.ac.be/cipl/bdlasla/](http://www.ulg.ac.be/cipl/bdlasla/)) of the ULg has created a large database of classical Latin texts (over 2,500,000 words). The database is currently still the only one that includes linguistic information, -the lemma and the complete morphological analysis of each unit of the text-, systematically checked by a philologist. The database is accessible either via the website *Latina Opera* (free access on [www.ulg.ac.be/CIPL/IsI.htm](http://www.ulg.ac.be/CIPL/IsI.htm); login and password necessary), or via the Hyperbase Web edition interface ([hyperbase.unice.fr](http://hyperbase.unice.fr)), developed as a joint research project with the UMR 7036 «Bases, corpus language» (CNRS-Université Côte d'Azur).

The LASLA is still expanding its database of classical and post-classical Latin texts. This paper will describe the structure of the information contained in the LASLA files, the reasons of the choices made by the LASLA in the way texts are lemmatised and tagged, the tools that the LASLA uses in order to lemmatise and tag its files (the LASLA Encoding Initiative interface and Collatinus-LASLA), the various LASLA corpora, the different ways to access the information contained in the LASLA files (*Opera Latina*, *Hyperbase Latin* and *Hyperbase Web Edition*), as well as the way the LASLA shares its data (the possibilities of Data Transfer Agreement with the ULiège). The paper will also present on the one hand the future projects of the LASLA and on the other hand its various collaborations with other research groups.

### **References.**

- DENOZ J., «*Opera latina*: une base de données sur internet». In *Euphrosyne*, 32, 2004: 79-88.
- DELATTE L., GOVAERTS S. & DENOZ J., *L'ordinateur et le latin. Techniques et méthodes, morphologie, syntaxe, lexicologie. Stylistique*, Liège, LASLA, 1978.
- ÉVRARD É., «Le laboratoire d'analyse statistique des langues anciennes de l'Université de Liège». In *Mouvement scientifique en Belgique*, 9, 1962: 163-169.
- LONGRÉE D. & POU DAT C., «New Ways of Lemmatizing and Tagging Classical and post-Classical Latin: the LATLEM project of the LASLA», in P. Anreiter & M. Kienpointner (ed.), *Proceedings of the 15th International Colloquium on Latin Linguistics*, (Innsbrucker Beiträge zur Sprachwissenschaft), Innsbruck, 2010, pp. 683-694.
- LONGRÉE D., PHILIPPART DE FOY C. & PURNELLE G., «Subordinate clause boundaries and word order in Latin: the contribution of the L.A.S.L.A. syntactic parser project LatSynt», in P. Anreiter & M. Kienpointner (ed.), *Proceedings of the 15th International Colloquium on Latin Linguistics*, (Innsbrucker Beiträge zur Sprachwissenschaft), Innsbruck, 2010, pp. 673-681.
- MELLET S. & PURNELLE G., «Les atouts multiples de la lemmatisation: l'exemple du latin». In MORIN A. & SÉBILLOT P., *JADT 2002, 6èmes Journées internationales d'Analyse statistique des Données Textuelles*. Saint-Malo: INRIA, IRISA, 2002: 529-538.