Philippe VERKERK and Yves OUVRARD

Collatinus: A New Tool for Latinists

Collatinus is a free open-source program to lemmatize Latin words[1]. It has been originally developed by Yves Ouvrard for teaching. It allows to generate a complete lexical aid, with a short translation and the morphological analyses of the forms, for any text which can be given to the students. The program knows the quantity of the vowels, which gives the possibility to scan a text. It also allows to search for a word in the dictionaries, either in a digital form or as images. It also includes a probabilistic tagger trained on the texts lemmatized by the LASLA[2]. As any open-source program[3], Collatinus can be tuned to meet any particular problem.

For any form, Collatinus tries to decompose it and analyze it. It then gives all the possible morphological analyses and root-words for the inflected form. To do that, Collatinus has a base of knowledge, which is easily expandable. It consists in a list of root-words (with a short translation, mainly in French and in English) and the rules for the flexion of each paradigms. All the files are in plain text and (almost) readable by the user. The automatic treatment of the digital dictionaries (L&S, Georges and Gaffiot 2016) allowed us to expand the lexical base from the original hand-made list of 11 000 root-words to a "complete" list of more than 80 000 words[4].

When vowel quantities are known for a given word, Collatinus can scan the form. When scanning a text, Collatinus applies the usual rules of lengthening, elision and hiatus. The metrical structure of verses can thus appear immediately. From the quantities, Collatinus can also deduce the position of the accent and it knows the basic rules for syllabification. The rhythm of the sentence (or of the *clausulae*) can thus be evidenced, which is important for studies in medieval Latin.

The examination of the texts lemmatized by the LASLA allowed us to retrieve statistical information about the use of the lemmas and the succession of three tags (based on the POS of the word and the morphological analysis). This information is used to order the solutions of the lemmatizations and to implement a probabilistic tagger based on a hidden 2nd order Markov model[5]. It is now possible to export in a CSV-format the results of the lemmatizer or of the tagger for further treatments.

The future developments can go in different directions. One possibility is to try to build the syntactic tree for any sentence: it will give more stringent conditions on the

---

[1] *"Collatinus, un outil polymorphe pour l'étude du latin"* by Y. Ouvrard and Ph. Verkerk, in Archivum Latinitatis Medii Aevi, 72, 305-311 (2014): http://documents.irevues.inist.fr/handle/2042/69526

[2] We thank D. Longrée and G. Purnelle who allow us to treat the texts lemmatized by the LASLA: http://web.philo.ulg.ac.be/lasla/textes-latins-traites

[3] Written in C++ (Qt 5). The application is designed with an intuitive GUI, but we also include a server (on an internal port of the computer), which allows a "command line" use of the functionalities. It is available for any platform (Mac OS X, Windows or Linux) on the website of the Equipex Biblissima: http://outils.biblissima.fr/en/collatinus. The sources are available: https://github.com/biblissima/collatinus

[4] This list should still be cleaned of some redundancies, but can be also complemented by words from other sources.

[5] See for instance "*A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*" by L. R. Rabiner, in Proceedings of the IEEE, 77, 257-285 (1989): http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/tutorial%20on%20hmm%20and%20applications.pdf

analyses of the forms. Another aim can be to expand the knowledge base of Collatinus toward Medieval Latin. One point is to handle the graphical variations in writing. But also, during its long life, the Latin language has considerably evolved. Some words have disappeared, others have appeared (for instance those coming from vernacular languages), many of them have changed their meaning, their morphological paradigm, their scansion, their spelling. In its last (beta) version, the classical data of Collatinus can be modified by the activation of lexical modules adapted to the corpus to be lemmatized. This activation can be manual, but also automatized by a pre-analysis.