Basic corpus annotation made easy: The Language Analysis Portal (LAP)

Jarle Ebeling, Emanuele Lapponi and Milen Kouylekov
(University of Oslo, Norway)

This poster presentation describes and showcases the CLARINO Language Analysis Portal (LAP, http://www.mn.uio.no/ifi/english/research/projects/clarino/), developed and maintained at the University of Oslo. LAP was launched in September 2016 after having been under development since 2013 (see Lapponi et al. 2013). Researchers interested in using the tools within LAP, can log in using their CLARIN or eduGAIN user account.

LAP is a user-friendly web interface to common annotation tasks such as tokenization, POS tagging and syntactic parsing of naturally occurring text. Basic annotation of texts is nearly always necessary as the first step in creating a corpus resource, whether it is intended to be a searchable resource or used as input for further analysis, e.g. semantic annotation. However, the tools available to achieve even the most basic annotation often demand non-trivial programming skills or installing and running different pieces of software one after the other from the command line. LAP helps the so-called "non-command-line-proficient researcher" to overcome these obstacles by letting the user drag, drop and link tasks in the web interface to create a workflow. The workflow can be stored and re-used on other data sets and shared with other users.

LAP is integrated with the Galaxy workflow system (https://galaxyproject.org/) to allow users to specify and run a series of tasks, known as a workflow, e.g. segmentation, tagging and parsing, and then download the output from all, or only the final of, these tasks. Originally developed for biology, Galaxy has been adapted to a range of other research areas, e.g. linguistics and social sciences. At the University of Oslo there is a team of developers devoted to maintaining and developing similar workflow systems for the life sciences, geology and metrological data in addition to linguistics and the social sciences.

LAP runs on a high-performance computer, handles large data sets, and recognises several languages, e.g. English, Norwegian and Sami. The members of the LAP team do not develop NLP processing tools themselves, but rather implement and make state-of-the-art (NLP) tools available to end users in an easy-to-use interface (Lapponi et al. 2015). The poster presentation will illustrate some of these tools in action.

The LAP development team welcomes suggestions for further open source tools to be integrated in the portal.

References

Lapponi, E., E. Velldal, N.A. Vazov and S. Oepen. (2013). Towards large-scale language analysis in the cloud. In *Proceedings of the workshop on Nordic language research infrastructure at NODALIDA 2013, May 22-24, 2013, Oslo, Norway*. [NEALT Proceedings Series 20]. Linköping University Electronic Press, 1–10. <http://www.ep.liu.se/ecp/089/001/ecp1389001.pdf>

Lapponi, E., S. Oepen, A. Skjærholt and E. Velldal. (2015). LAP: The CLARINO Language Analysis Portal. In *CLARIN Annual Conference 2015: Book of Abstracts*, *Wroclaw, Poland, October 14-16, 2015*, 43–48. <https://www.clarin.eu/sites/default/files/book%20of%20abstracts%202015.pdf>