LAS: an integrated language analysis tool for multiple languages

Eetu Mäkelä[1]

[1]Aalto University

29 June 2016

Paper DOI: http://dx.doi.org/10.21105/joss.00035
Software Repository: https://github.com/jiemakel/las
Software Archive: https://dx.doi.org/10.5281/zenodo.160256

Summary

LAS is a command-line tool for lemmatizing, morphological analysis, inflected form generation, hyphenation and language identification of multiple languages.

These functionalities are of use as part of many workflows requiring natural language processing. Indeed, LAS has been used for example as part of a pipeline for entity recognition (Mäkelä 2014), in creating a contextual reader for texts in English, Finnish and Latin (Mäkelä, Lindquist, and Hyvönen 2016), and for processing a Finnish historical newspaper collection in preparation for data publication (Pääkkönen et al. 2016).

The functionalities of LAS are mostly based on integrating existing tools into a common package. Particularly, the tool bases on: * Finite state transducers provided by the HFST (Linden et al. 2013), Omorfi (T. A. Pirinen 2015) and Giellatekno (Moshagen et al., n.d.) projects * Snowball stemmers * the language-detector library * Statistical language models from Turku NLP (Haverinen et al. 2014)

While LAS supports many languages, the most complete support it has is for Finnish, where considerable work has gone into improving the results.

Aside from a being available as a command-line tool, the functionalities in LAS are also available as a web service, at http://demo.seco.tkk.fi/las/.

References

Haverinen, Katri, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski, and Filip Ginter. 2014. "Building the Essential Resources for Finnish: The Turku Dependency Treebank." *Language Resources and Evaluation* 48 (3). Springer Netherlands: 493–531. doi:10.1007/s10579-013-9244-1.

Linden, Krister, Erik Axelson, Senka Drobac, Sam Hardwick, Juha Kuokkala, Jyrki Niemi, Tommi Pirinen, and Miikka Silfverberg. 2013. "HFST – a System for Creating NLP Tools." In *Systems and Frameworks for Computational Morphology*, edited by Cerstin Mahlow and Michael Piotrowski, 380:53–71. Communications in Computer and Information Science. Berlin Heidelberg: Springer.

Mäkelä, Eetu. 2014. "Combining a REST Lexical Analysis Web Service with SPARQL for Mashup Semantic Annotation from Text." In *The Semantic Web: ESWC 2014 Satellite Events: ESWC 2014 Satellite Events, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers*, edited by Valentina Presutti, Eva

Blomqvist, Raphael Troncy, Harald Sack, Ioannis Papadakis, and Anna Tordai, 424–28. Cham: Springer International Publishing. doi:10.1007/978-3-319-11955-7_60.

Mäkelä, Eetu, Thea Lindquist, and Eero Hyvönen. 2016. "CORE - a Contextual Reader Based on Linked Data." In *Proceedings of Digital Humanities 2016, Long Papers*. Kraków, Poland.

Moshagen, Sjur, Jack Rueter, Tommi Pirinen, Trond Trosterud, and Francis M Tyers. n.d. "Open-Source Infrastructures for Collaborative Work on Under-Resourced Languages." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC*, 71–77.

Pääkkönen, Tuula, Jukka Kervinen, Asko Nivala, Kimmo Kettunen, and Eetu Mäkelä. 2016. "Exporting Finnish Digitized Historical Newspaper Contents for Offline Use." *D-Lib Magazine* 22 (7/8). Corporation for National Research Initiatives. doi:10.1045/july2016-paakkonen.

Pirinen, Tommi A. 2015. "Omorfi—Free and Open Source Morphological Lexical Database for Finnish." In *Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania*, 313–15. 109. Linköping University Electronic Press.