LAS: an integrated language analysis tool for multiple languages

Eetu Mäkelä

Aalto University

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 (Lindén 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 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


Mäkelä, Eetu, Thea Lindquist, and Eero Hyvönen. 2016. “CORE - a Contextual Reader...

