Alexander MEHLER and Tim GEELHAAR

Frankfurt Latin Lexicon: From Morphological Expansion to Latin Word Embeddings and Lexical Networks

We present the Frankfurt Latin Lexicon (FLL) as a lexical resource used by us in a number of NLP tasks of preprocessing Latin texts such as morphological tagging, lemmatization, and POS tagging. FLL was developed with the help of several source lexicons and taggers. First, a large number of so-called superlemmas were collected, then variants (lemmata) were differentiated for each superlemma, and finally a rule-based morphological expansion was carried out for each of these lemmas. The resulting lexicon is used for hu- man computation, according to which its entries are continuously checked and, if necessary, corrected by registered expert users. FLL serves as a reference lexicon of TEILex, a system for integrating lexica and text corpora, in which the tokens of a corpus are linked with their lexicon entries in such a way that updates of the lexicon are immediately transferred to the linked corpora and vice versa. In this way, expert-based lexicon modeling becomes independent of indexing the underlying corpus. The paper describes the use of FLL and TEILex as two text-technological resources in different tasks such as morphological tagging, lemmatization and POS tagging as well as in the calculation of so-called Wikiditions of Latin corpora and their distribution via the website CompHistSem (http://www.comphistsem.org/home.html). Furthermore, the extension of FLL by various methods for the calculation of word embeddings is presented and illustrated by lexical networks and their analysis for the purpose of text classification. In this sense, the paper spans a spectrum from re- source development (FLL) and NLP (tagging and lemmatization) to text mining (word embeddings and lexical networks) and the provision of research infrastructures and NLP pipelines (TextImager, eHumanities Desk- top and CompHistSem).