HunLex - a framework for morphological dictionaries

Viktor Trón*

In this article, we present HunLex, a morphological resource-specification framework and resource compiler tool which is being developed as part of the Budapest Institute of Technology Media Education and Research Center's HunTools NLP toolkit (see `http://www.szoszablya.hu`).

HunLex offers a formalism for specifying a base lexicon and morphological rules which can then serve as a central database capable of providing language-specific knowledge to a variety of NLP tools. The prototype implemented for the Szószablya project at the BIT is able to provide optimized lexical resources for the HunTools MorphBase routines (spell-checker, stemmer, morphological analyzer/generator). These resources are compiled offline from the central lexicon and grammar in a highly configurable way so that users can fine-tune these resources according to their needs.

The motivation behind HunLex came from two opposing types of requirements lexical resources are supposed to fulfill: (i) scalability, maintainability, extensibility; and (ii) optimized format for the application. The constraints in (i) favour one central, redundancy-free, abstract, but transparent specification, while (ii) requires various application-specific, and potentially redundant, optimized formats. In order to reconcile these requirements, HunLex introduces an offline layer which mediates between the two levels of resources: a central database conforming to (i), which is ideal for human maintanance, and the various specific formats that are inputs to software modules conforming to (ii) for performance. HunLex is used to compile the base resources into an application-specific format (called dic and aff files in the case of the MorphBase routines) in a configurable way. This includes the choice of format for algorithm (spell-checking, stemming, morphological analysis or generation), selection of morphemes, grouping of morphemes to be stripped as a cluster (with one rule application), selection of morphophonological features that are to be observed or ignored, depth of recursive rule application, selection of registers and degree of normativity based on usage qualifiers in the database.

The HunLex framework is used in the development of an open-source morphological database (lexicon and grammar) for the Hungarian language in a collaboration between the Research Institute for Linguistics and the MERC Lab, which aspires to be the most complete and accurate account of Hungarian morphology published so far.

* International Graduate College, Saarland University and University of Edinburgh, `v.tron@ed.ac.uk`