Open source morphological analyzer

Németh László*, Halácsy Péter* Kornai András**, Trón Viktor***

The HunTools natural language processing toolkit emerged from the SzolSzablya morphological analyzer project at the the Budapest Institute of Technology Media Education and Research Center. In this paper we concentrate on the architecture of the MorphBase morphological component which supports spellchecking, stemming, morphological analysis, and generation in a set of language-neutral routines, and describe the Hungarian-specific resources. Both the Hungarian-specific and the language-neutral parts of the system are available under the open source LGPL license.

The core engine of MorphBase is an extension of the well-known open-source ispell spellchecker, which identifies correctly spelled words by first stripping affixes according to a rule-system and then looking up the stems from a lexicon. Both the rule-system and the lexicon are specified as input files and compiled off-line. Our improved version is similarly language independent (and compatible with ispell file formats) but has significant additional functionality. First, we enabled the output of stripped forms, thereby creating a stemmer. Next, we enabled alternative analyses, both in stemming and providing a full morphological analysis. Finally, we replaced the simple the simple one-pass affix stripping mechanism of ispell by a recursive system in which affixes can be stripped in as many layers as needed. This results in considerable simplification of the lexical resource, as well as increased linguistic transparency and maintainability.

* Budapest University of Technology Centre for Media Research and Education, {nemeth,halacsy}@mokk.bme.hu

** MetaCarta Inc., andras@kornai.com

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