VIRGINIJUS DADURKEVIČIUS

LITHUANIAN MORPHOLOGY IN THE *HUNSPELL* FRAMEWORK

Summary

The paper presents the results of an attempt to build basic Lithuanian language resources using the widespread *Hunspell* platform. The spelling is actually the primary target of this open-source platform but the morphological analysis and synthesis are also possible. Moreover, the ability to efficiently perform lemmatization (stemming) makes this platform the best option for text search engines (e.g. *Solr/Lucene*) and information retrieval. Taggers, grammar checkers and other basic natural language processing tools can also be build using properly built *Hunspell* language resources.

Every *Hunspell* language resource consists of two files: dictionary and affixes (it may be empty). The dictionary contains main forms (lemmas) whereas the affixes contain the morphological rules to generate all possible forms. As a source for the dictionary we have used the Modern Lithuanian Dictionary (6-th edition), Corpus of the Contemporary Lithuanian Language compiled at the Center of Computation Linguistics of Vytautas Magnus University, database of documents of the Lithuanian Parliament, *versti.eu* machine translation corpus of Vilnius University and various public internet sources (totally 1.3 billion tokens). Main criteria for semi-manual compilation of the Lithuanian dictionary of lemmas from these sources was correctness, usability, actuality and approval by language authorities. Deprecated loanwords or extremely rare, exotic, obsolete, jargon, insulting forms were discarded from the list. Resulting dictionary consists of 171 000 lemmas: 42 000 common nouns, 73 000 proper nouns, 15 000 adjectives, 53 pronouns, 153 numerals, 35 000 verbs, 4 000 adverbs and 2 000 others (prepositions, conjunctions, particles, onomatopoeias, interjections, acronyms and abbreviations).

The second component of language resource, the so called “affix file”, contains information of various kind: metadata, preferable suggestions for spelling correction, grouping of rules, explicit tags for flexing and non-flexing properties, rules for suffix and affix alteration.

In order to make the *Hunspell* resources suitable for creating basic language tools, e.g. morphological analyzer and synthesizer, some principles should be kept:

1) every flexion paradigm (consisting of one or more rules) should be thoroughly generated from one single lemma in dictionary file (it is not trivial, especially for irregular verbs);

2) every individual alteration case should have its own morphological tag, e.g. ‘*Masc_Sg_Il*’ for masculine + singular + illative;
3) every dictionary item should have references for part of speech and other non-flexing information;
4) avoid prefixation via rules, use dictionary instead – affixed forms may have completely different meanings and using them under single lemma may cause problems for text search engines;
5) do not rely much on calling rules from rules – calling depth can by no more than 1.

The coverage of the contemporary Lithuanian by this implementation of Lithuanian morphology is about 98 percent. The full list of all the theoretically possible forms generated by this resource contains about 17 million entries.

This work clearly shows an efficient way for any language (especially with scarce funding resources) to make basic language tools using a single open source development platform – the *Hunspell*.

KEYWORDS: Lithuanian, grammar, morphology, computational linguistics, Hunspell, speller, tagger, analyser, parsing, disambiguation, indexing, search.

VIRGINIJUS DADURKEVIČIUS

Institute of Applied Research
Vilnius University
M. K. Čiurlionio g. 29, 03100 Vilnius, Lithuania
virginijus.dadurkevicius@tmi.vu.lt
dadurka@gmail.com