The Szószablya project – www.szoszablya.hu

Péter Halácsy[1], András Kornai[2], László Németh[1], András Rung[3], István Szakadát[1], and Viktor Trón[4]

[1] Centre of Media Research and Education, Budapest University of Technology and Economics, {halacsy,szakadat,rung}@mokk.bme.hu
[2] MetaCarta Inc., andras@kornai.com
[3] Center of Cognitive Science, Budapest University of Technology and Economics, rung@itm.bme.hu
[4] School of Informatics, University of Edinburgh, v.tron@ed.ac.uk

The goal of Szószablya (Wordsword) project[5] is to create the *Hungarian Webcorpus*, the *Szószablya Frequency Lexicon*, *hunmorph*, an LGPL licensed morphological analyzer, *hunstem*, a stemmer and *hunspell*, a spellchecker and *hunlex* a Hungarian spelling and morphological dictionary which is utilized by the other applications. The project has been launched in March 2003.

The Hungarian Webcorpus is a tokenized collection of Hungarian language texts which is significantly more comprehensive than the previously existing ones: it is based on a collection of 2.4 million web pages, which after basic distilling gave rise to a webcorpus with 670 million tokens and 15 million token types. The documents were collected from the .hu domain, in December of 2002, using the Larbin webcrawler. The pages have been normalized and tokenized into words and sentences. Using automated data cleaning techniques 433,000 good quality web pages were then selected (113 million tokens, 4.5 million token types). In December of 2003 a much larger corpus and the corresponding frequency lexicon will be made available.

Based on raw and selected corpora, two versions of word frequency list have been produced, indicating both token frequency and document frequency. The 4 (in raw lexicon) and 2.8 (in selected lexicon) million words which were judged as correct Hungarian words by the hunspell spellchecker's current version were selected. Unrecognized words form the basis for a large-scale extension of the hunlex dictionary using machine-supported human annotation. According to our estimation, the current version of hunspell recognizes at least 96% of the correctly spelled words on Hungarian web pages.

Hunspell is the modified version of the Magyar MySpell program developed by László Németh and have been available since the early stages of the project. The applications hunmorph and hunstem, the functional design of which is in progress, are based on hunspell's architecture and sourcecode. The first versions are to be released in May 2004 as the project finishes. An important feature of hunmorph is that – unlike the spellchecker's strict adherence to norms — it is also capable of analyzing forms deviant in terms of morphology and/or spelling.

[5] Managed by Centre of Media Research and Education of Budapest University of Technology and Economics, supported by Ministry of Informatics and Communication, origo.hu and MATÁV Rt.