Documentation to compiling a list of Hungarian roots including parsing and transcription

Ildikó Emese Szabó

July 1st, 2016

1 Raw materials

We used the Szótár AdatBázis 1.0 (Szilágyi N., 2014) as a basic corpus, Siptár and Törkenczy (2000) for 'shortening' stems (stems exhibiting stem vowel shortening) and Rung (2012) for epenthetic stems. Since the Szótár AdatBázis 1.0 is originally a dictionary, several words occur as multiple lexical entries (e.g. *nyom*^{1} 'push$_1$' and *nyom*^{2} 'trace$_N$'), first a list had to be created only including every word once and excluding hyphenated and space-separated entries as those are surely polymorphemic. This list consists of 71137 word forms (henceforth: wordlist).

2 Created resources

Four lists of affixes were manually compiled (preverbs, prefixes, derivational suffixes, inflectional suffixes), containing all allomorphs of the appropriate morphemes. Shortening stems, epenthetic stems were manually collected into a separate list for bound stems. There are two completely predictable stem alternation processes in Hungarian: Low Vowel Lengthening (*kutya* 'dog', *kutyá-k* 'dogs'; *kefe* 'brush', *kefé-k* 'brushes') and the vowel $\sim \emptyset$ of words ending in *-alom, -elem* $\sim$ *-alm-, -elm-*. A list for these completely predictable bound stems was automatically generated based on the appropriate words in wordlist. Bi- and trimorphemic words were filtered out. Additionally, a prefix banlist was identified, containing "prefixes" (e.g. *mikro, makro, pre*) that only or mostly occur in foreign loanwords. The same list was created for endings and suffixes.

3 Selection

In order to make sure that our results are not influenced by more variation of vowel combinations in loanwords, any word beginning with a banned prefix or ending in a banned "suffix" was filtered out. After filtering out polymorphemic and foreign forms, 17705 roots remained. There were 5 words in Rung (2012) that were not in Szilágyi N. (2014), those 5 were added to our corpus. As a result of hapax legomena and proper names as parts of compounds, these 17710 words contained some which were still polymorphemic. These words could be

identified based on length. Words longer than 9 characters were manually searched through to identify further 3428 words which were polymorphemic, but the morphemes did not appear in either the initial corpus or our lists. These were later discarded, leaving a total of 14282 roots.

References

Rung, András. 2012. A magyar főnévi alaktani jelenségek analógiás megközelítésben [An analogical aproach to phenomena in hungarian noun morphology]. Doctoral Dissertation, Eötvös Loránd Tudományegyetem.

Siptár, Péter, and Miklós Törkenczy. 2000. *The Phonology of Hungarian*. The Phonology of the Worlds languages. New York: Oxford University Press.

Szilágyi N., Sándor. 2014. Szótár AdatBázis 1.0, Property of Babeş-Bolyai University, Cluj-Napoca, Romania. Property of Babeş-Bolyai University, Cluj-Napoca, Romania.