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$_V$' 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

Lists of shortening stems, epenthetic stems were manually collected into a separate list for bound stems. For epenthetic stems, we used the Appendix of Rung (2012) as a basis. However, those were often compounds. To filter these out, we used the process we explain in the *Section 3*, which led to 126 stems. These were overlooked manually resulting in a final list of 112 monomorphemic epenthetic stems.

There are two entirely 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.

Four lists of affixes were manually compiled (preverbs, prefixes, derivational suffixes, inflectional suffixes), containing all allomorphs of the appropriate morphemes. 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 (e.g. *mikro-*), ending in a

banned "suffix" (e.g. *-lógia*) or containing the letter *'x'*, which exclusively appears in foreign loanwords were thrown away. After filtering out polymorphemic and foreign forms, 8805 roots remained. There were 5 words in Rung (2012) that were not in Szilágyi N. (2014),[1], they were manually added to our corpus. As a result of hapax legomena and proper names as parts of compounds, these 8810 words contained some which were still polymorphemic. These words could be identified based on length. Words longer than 7 characters were manually searched through to identify further 1925 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 6885 roots.

4 For internal use

Ancillary files:

1. preverbs
2. prefixes
3. derivational suffixes
4. inflectional suffixes

For 3 and 4 I only included the ones we needed for this (e.g. infinitival *-ni* and possessive forms, case-markers, etc are not included in the original dictionary, so I didn't include them either)

5. automatically generated bound stems: I looked word-final a's and e's and replaced them with á's and é's. Similarly: *-alom* → *-alm*; *elem* → *-elm*. these are the exceptionless bound stems.
6. a "misc" of bound stems: stem-vowel shortening: based on lists from the literature. v-adding stems: I only found references to a book I couldn't get (The sound pattern of Hungarian, Vágó, 1980) and a total number (19), so I came up with them. This has to be checked with Vágó (1980), though.
7. epenthetic bound stems: I got the main list from the Appendix of Rung (2012), which is a PhD dissertation in Hungarian. However, those were often compounds, I just used the parser I've been building to throw those out. That left me with 126 stems, I just looked through those manually and threw out 14 polymorphemic ones.

[1] *boholy* 'fluff', *cseber* 'a kind of bucket', *kölök* 'kid, kiddo', *pöcök* 'a small lever', *veder* 'bucket'

5 & 6 & 7: I checked that there is no overlap between them. There were 5 words in 7 that were not in our dictionary (mostly really infrequent), I added those 5, but I can take them out if necessary.

8. a banlist for prefixes
9. a banlist for suffixes

what 8. and 9. do is they contain suffixes (&endings) and prefixes that can be used to identify foreign words. But practically, it does not only contain loan-word endings: I added frequent combinations of derivational suffixes (since the parser can only deal with trimorphemic things at best) and endings like *X-féle* 'a kind of X' and *X-szerű* 'X-like', which often take proper names as X. This is the part that I feel less certain about methodologically, but I go through the lists that they discard and choose the endings to include based on the ending-ordered dictionary and the losses are negligible – especially compared to the efficiency.

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.