Morphological Modifications in Szeged Corpus 2.5

Veronika Vincze[1], Viktor Varga[2], Katalin Ilona Simkó[2],
János Zsibrita[2], Ágoston Nagy[2], Richárd Farkas[2]

[1]Hungarian Academy of Sciences, Research Group on Artificial Intelligence
[2]University of Szeged, Department of Informatics
{vinczev,zsibrita,nagyagoston,rfarkas}@inf.u-szeged.hu
{viktor.varga.1991,kata.simko}@gmail.com

In this work, we present Szeged Corpus 2.5, in which we applied some morphological modifications which we believe will benefit real-world NLP applications. The modifications involve the introduction of new codes in the coding system as well as the correction of some morphological codes, with special emphasis on misspelled words.

Recently, there has been a successful attempt to harmonize the coding systems MSD and KR. The two coding systems cannot be mapped in a one-to-one way, so if we want to exploit both resources in a statistical language parser (POS tagger, constituency parser, dependency parser etc.), we have to employ conversion rules, which leads to the loss of information. In order to prevent this, the two coding systems (MSD and KR) were harmonized and their basic principles were also made compatible.

Here, we applied the principles of the harmonized morphology in the annotation of Szeged Corpus 2.5. For instance, only those pieces of derivational information are explicitly marked that are expressed with syntactic tools in other languages. We applied this approach to verbs with frequentative, modal and causative suffixes and the lemma became the word form without any of the above mentioned suffixes.

As for the treatment of adverbial pronouns, we decide to derive them from personal pronouns and thus inserted them into the pronominal system of morphological codes.

Present, past and future participles were also given a new code since in the earlier version of the corpus, they could not be distinguished on the basis of their codes, what is more, their code coincided with that of adjectives.

We also eliminated the differentiation between proper nouns and common nouns at the level of morphology. In addition to the morphological modifications described above, we also paid attention to the correction of misspelled words. All in all, changes involved about 4.36% of the tokens in the corpus.

These modifications also made it possible to train and evaluate the morphological analyzer and POS-tagger modules of magyarlanc on the new version of the corpus. According to our results, the accuracy of POS-tagging does not change significantly as compared to that achieved by training magyarlanc on Szeged Corpus 2.0.