Morphological and Syntactic Annotation of Hungarian Webtext

Veronika Vincze[1,2], Viktor Varga[1], Petra Anna Papp[1],
Katalin Ilona Simkó[1], János Zsibrita[1], Richárd Farkas[1]

[1]University of Szeged, Department of Informatics
Szeged, Árpád tér 2.
{vinczev,zsibrita,rfarkas}@inf.u-szeged.hu,
{varga.viktor.1991,papp.petra.anna,kata.simko}@gmail.com
[2]MTA-SZTE Research Group on Artificial Intelligence
Szeged, Tisza Lajos körút 103.

For a while now, internet communication has been used as a source of data for research. Texts on the web trying to mimic oral communication include many abbreviations and errors that make their linguistic processing more difficult. Our goal was to create a corpus of texts from the web and manually annotate it for morphology and syntax in order to make it useful for the development of future natural language processing applications for this domain.

Our corpus is made up of public Facebook comments (1208 sentences, 8615 tokens) and questions and answers from gyakorikerdesek.hu (728 sentences, 9702 tokens). Most posts are about users' hobbies, personal interests and lifestyle.

First, we manually segmented the sentences and tokenised the text, then, using one of the modules of magyarlanc, we built a corpus, structurally similar to the Szeged Korpusz, in which the annotators manually assigned the contextually correct morphological code to each word. Similar to Szeged Treebank and Szeged Dependency Treebank, we also created manual constituent and dependency syntax analysis for each sentence. We mainly followed the principles used in the development of our two previous, bigger treebanks, but some modifications were unavoidable given the special form of this text. The corpus is also annotated for semantic and discourse level uncertainty markers and we plan to annotate named entities in it as well.

This first Hungarian, manually annotated web corpus will be used as a test database in developing a morphological and syntactic parser, optimalised for the analysis of texts from the web. The corpus is currently too small to train statistical parsers, however, our goal was to create a benchmark database. We believe that as web texts are so varied both in topic and genre, the application of supervised machine learning techniques would not be a suitable solution, instead, we plan to use domain adaptation methods.