Manually Annotated Hungarian Natural Language Corpus: the Szeged Korpusz

Dóra Csendes[1], Csaba Hatvani[1],
Zoltán Alexin[1], János Csirik[1], Tibor Gyimóthy[1], Gábor Prószéky[2], Tamás Váradi[3]

[1] University of Szeged, Department of Informatics
H-6720 Szeged, Árpád tér 2., Hungary
{dcsendes, hacso, alexin, csirik, gyimi}@inf.u-szeged.hu
http://www.inf.u-szeged.hu
[2] MorphoLogic Ltd. Budapest
H-1118 Budapest, Késmárki u. 8., Hungary
proszeky@morphologic.hu
http://www.morphologic.hu
[3] Research Institute for Linguistics at the Hungarian Academy of Sciences
H-1068 Budapest, Benczúr u. 33., Hungary
varadi@nytud.hu
http://www.nytud.hu

Keywords: natural language corpus, morpho-syntactic annotation, shallow parsing, treebank building

The present state of the Szeged Korpusz is the result of three national projects and the cooperation of the following three consortium partners: the University of Szeged, Department of Informatics, MorphoLogic Ltd. Budapest, and the Research Institute for Linguistics at the Hungarian Academy of Sciences.

The corpus currently comprises 1.2 million words plus 290 thousand punctuation marks. Texts have gone through different phases of natural language processing and analysis. First, they were segmented into manageable units, i.e. words, punctuation marks and special tokens.

In the second step, corpus words were morpho-syntactically analysed with the help of an automatic analyser (the HuMor Hungarian morphological tagger developed by MorphoLogic Ltd.) and then manually POS tagged by linguistic experts.

In the following phase of processing, texts of the Szeged Korpusz have been shallow parsed. During this phase annotators marked noun phrase structures and the clause structure of corpus sentences.

Current works aim at a more detailed syntactic analysis of the texts including the annotation of adverbial, postpositional, adjectival structures and the identification of verbs and their argument structures. With this, we intend to lay the foundation of a Hungarian treebank that – as a continuation of the work – is planned to be enriched with semantic information as well.

The Szeged Korpusz is publicly available after on-line registration (http://www.inf.u-szeged.hu/lll) and can be used free of charge for educational and research purposes.