The Czech Academic Corpus version 1.0 has been released

Barbora Vidová Hladká

The Czech Academic Corpus version 1.0[1] is a corpus with a manual annotation of morphology of the Czech language consisting of approximately 600,000 words in continuous texts.

The process and the size of the Czech Academic Corpus (CAC) project differs from every traditional project. The primary goal of this project was to create a computerized corpus that would contain manual annotation of morphology and syntax of Czech. This manual annotation of morphology and syntax was initially developed more than twenty years ago (1971-1985) at the Institute of the Czech Language of the Czech Academy of Science as a basis for constructing a frequency dictionary of Czech at that time.

Having mentioned the years when the history of CAC has started we cannot miss the fact that there were available two computerized annotated corpora in the 1960s - Brown Corpus of American English and LOB Corpus of British English. Both corpora became well known to the corpus linguists whereas CAC (although containing richer annotation schemes) has remained hidden mainly because of the political regime of the 1980s in the Czech Republic. Fortunately for Czech computational linguistics, CAC was (and still is, obviously) of significant importance because the very first experiments on the corpus-based processing of Czech could be performed thanks to CAC.

Independent from CAC, the project of the Prague Dependency Treebank (PDT) was launched in 1996. The complexity of the PDT three-layer annotations (morphological, syntactic-analytical, tectogrammatical) is reflected in the volume of the annotated data - 2 million words are annotated only morphologically, 1,5 million words have also an analytic annotation and from them 800,000 words is provided with tectogrammatical annotation. The second version of PDT has been released in summer 2006 (Hajič et al., 2006), (PDT 2.0, 2006). The experience acquired from annotating such a huge volume of data is so exceptional and illuminating that it has become one of the main motivations for further work within CAC - the idea of converting the internal format and annotation schemes of CAC in a way that they would be compatible within PDT was proposed. This conversion will facilitate the possibility of integrating CAC annotations directly into PDT.

The currently released first version of CAC (CAC 1.0, 2006) is a result of conversion of the internal format and morphological annotations. The guide to the CAC 1.0 (Hladká et al., 2006) is a road map to the CD-ROM which offers:

- data material (CAC 1.0) for theoretical linguists (that reflects real language usage) and for computational linguists (that should contribute to the amelioration of the applications of natural language processing that cannot exist without analysis of the texts on the morphological level),
- a graphic tool (Bonito) enabling searching and viewing CAC 1.0,
- a lexical annotation workbench (LAW) - an environment for browsing and annotation of data,
- applications for morphological analysis and tagging,
- an electronic exercise book of Czech (STYX) - interesting teaching aid for teachers and students that can be used in Czech lessons for practice in Czech morphology.

[1] The Czech Academic Corpus v. 1.0 is published within the frame of the project "Resources and Tools for Information Systems" No. 1ET101120413 financed by the Grant Agency of the Academy of Sciences of the Czech Republic.

The CAC project continues with conversions of syntactic annotations (Ribarov and Hladká, 2006) into the chosen concept in PDT that should result into the edition of the second version. This second version will facilitate the direct integration of CAC into PDT.

A reader is more than welcome to visit the webpage `http://ufal.mff.cuni.cz/rest` and order CAC 1.0.

References

CAC 1.0. 2006. http://ufal.mff.cuni.cz/rest.

Hajič, Jan, Eva Hajičová, Jarmila Panevová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, and Marie Mikulová. 2006. Prague Dependency Treebank 2.0 - CD ROM, ISBN:1-58563-370-4. Linguistic Data Consortium.

Hladká, Barbora, Jan Hajič, Jiří Hana, Jaroslava Hlaváčová, Jiří Mírovský, and Jan Votrubec. 2006. *Czech Academic Corpus 1.0 Guide*. Karolinum - Charles University Press.

PDT 2.0. 2006. http://ufal.mff.cuni.cz/pdt2.0.

Ribarov, Kiril, Alla Bémová and Barbora Hladká. 2006. When a statistically oriented parser was more efficient than a linguist: A case of treebank conversion. *Prague Bulletin of Mathematical Linguistics*, 86.