Prague Dependency Treebank as an exercise book of Czech

Barbora Hladká and **Ondřej Kučera**
Institute of Formal and Applied Linguistics
Charles University
Malostranské nám. 25
118 00 Prague, Czech Republic
hladka@ufal.mff.cuni.cz, ondrej.kucera@centrum.cz

Abstract

There was simply linguistics at the beginning. During the years, linguistics has been accompanied by various attributes. For example *corpus* one. While a name corpus is relatively young in linguistics, its content related to a language - collection of texts and speeches - is nothing new at all. Speaking about corpus linguistics nowadays, we keep in mind collecting of language resources in an electronic form. There is one more attribute that computers together with mathematics bring into linguistics - *computational*. The progress from working with corpus towards the computational approach is determined by the fact that electronic data with the ”unlimited” computer potential give opportunities to solve natural language processing issues in a fast way (with regard to the possibilities of human being) on a statistically significant amount of data.

Listing the attributes, we have to stop for a while by the notion of *annotated* corpora. Let us build a big corpus including all Czech text data available in an electronic form and look at it as a sequence of characters with the space having dominating status – a separator of words. It is very easy to compare two words (as strings), to calculate how many times these two words appear next to each other in a corpus, how many times they appear separately and so on. Even more, it is possible to do it for every language (more or less). This kind of calculations is language independent – it is not restricted by the knowledge of language, its morphology, its syntax. However, if we want to solve more complex language tasks such as machine translation we cannot do it without deep knowledge of language. Thus, we have to transform language knowledge into an electronic form as well, i.e. we have to formalize it and then assign it to words (e.g., in case of morphology), or to sentences (e.g., in case of syntax). A corpus with additional information is called an annotated corpus.

We are lucky. There is a real annotated corpus of Czech – Prague Dependency Treebank (PDT). PDT belongs to the top of the world corpus linguistics and its second edition is ready to be officially published (for the first release see (Hajič et al., 2001)). PDT was born in *Prague* and had arisen from the tradition of the successful Prague School of Linguistics. The *dependency* approach to a syntactical analysis with the main role of verb has been applied. The annotations go from the morphological level to the tectogrammatical level (level of underlying syntactic structure) through the intermediate syntactical-analytical level. The data (2 mil. words) have been annotated in the same direction, i.e., from a more simple level to a more

complex one. This fact corresponds to the amount of data annotated on a particular level. The largest number of words have been annotated morphologically (2 mil. words) and the lowest number of words tectogramatically (0.8 mil. words). In other words, 0.8 million words have been annotated on all three levels, 1.5 mil. words on both morphological and syntactical level and 2 mil. words on the lowest morphological level.

Besides the verification of 'pre-PDT' theories and formulation of new ones, PDT serves as training data for machine learning methods. Here, we present a system **Styx** that is designed to be an exercise book of Czech morphology and syntax with exercises directly selected from PDT. The schoolchildren can use a computer to write, to draw, to play games, to page encyclopedia, to compose music - why they could not use it to parse a sentence, to determine gender, number, case, . . . ? While the Styx development, two main phases have been passed:

1. **transformation** of an academic version of PDT into a school one. 20 thousand sentences were automatically selected out of 80 thousand sentences morphologically and syntactically annotated. The complexity of selected sentences exactly corresponds to the complexity of sentences exercised in the current textbooks of Czech. A syntactically annotated sentence in PDT is represented as a tree with the same number of nodes as is the number of the words in the given sentence. It differs from the schemes used at schools (Grepl and Karlík, 1998). On the other side, the linear structure of PDT morphological annotations was taken as it is – only morphological categories relevant to school syllabuses were preserved.
2. **proposal** and **implementation of exercises**. The general computer facilities of basic and secondary schools were taken into account while choosing a potential programming language to use. The Styx is implemented in Java that meets our main requirements – platform-independent system and system stability.

At least to our knowledge, there is no such system for any language corpus that makes the schoolchildren familiar with an academic product. At the same time, our system represents a challenge and an opportunity for the academicians to popularize a field devoted to the natural language processing with promising future.

A number of electronic exercises of Czech morphology and syntax were created. However, they were built manually, i.e. authors selected sentences either from their minds or randomly from books, newspapers. Then they analyzed them manually. In a given manner, there is no chance to build an exercise system that reflects a real usage of language in such amount the Styx system fully offers.

References

Jan Hajič, Eva Hajičová, Barbora Hladká, Petr Pajas, Jarmila Panevová, and Petr Sgall. 2001. *Prague Dependency Treebank 1.0 (Final Production Label)* CD-ROM, CAT: LDC2001T10, ISBN 1-58563-212-0, Linguistic Data Consortium.

Miroslav Grepl and Petr Karlík 1998. *Skladba češiny. [Czech Langauge.]* Votobia, Praha.