

Barbora Vidová Hladká

research profile

November 2008

My background is in Computer Science, with graduate education in Computational Linguistics. Currently, I am a senior research associate in the **Institute of Formal and Applied Linguistics (UFAL)** at the Computer Science School of the Faculty of Mathematics and Physics, Charles University in Prague (CUNI).

I actively participate in formation of the modern Czech computational linguistics research since its very beginning. A corpus-based methodology has been applied to Czech language processing for the first time in the mid-1990s - namely to the morphological tagging that became a topic of my Master thesis, defended in 1994. The **Czech Academic Corpus (CAC)**, morphologically and syntactically annotated corpus built in 1970s-1980s in the Institute of Czech Language, Academy of Science of the Czech Republic, was of great importance, because it could be used as training data for the very first experiments. My tagging experiments turned out to be of crucial significance for the development of Czech language processing. From this point of view, the CAC essentially has influenced my research development.

In my PhD thesis, I continued in my research corpus-based methods to tagging. All the largely used tagging procedures were originally designed for English. These procedures are corpus-based and, in principle, language independent. So it is a challenge to apply them to any other language, if an annotated corpus for the given language exists. In 1997 at the Automatic Natural Language Processing Conference in Washington, USA, I presented a comprehensive evaluation of the statistical methodology applied to two typologically different languages – Czech and English – for the first time ever; no other comparison has been done before. A year before, in 1996, **I was awarded the Josef Hlavka's award** for the best students of CUNI for my research on tagging.

A project of the **Prague Dependency Treebank (PDT)** for manual annotation of a substantial amount of Czech-language data with linguistically rich information has started in 1996. I have been involved in this project since its beginning, namely I supervised a team of five co-workers in the morphological annotation. A project of PDT presents an exceptional achievement of the Czech computational linguistics because it pursues a systematic conceptual framework of the text annotation going from morphology to semantics through syntax. The family of the Prague dependency treebanks including not only PDT and CAC is being comparable only with a family of the Penn Treebanks of the University of Pennsylvania. Halfway through, in 1998, PDT version 0.5 was used as the main data set for the project in the prestigious **NSF Workshop: Language Engineering for Students and Professionals Integrating Research and Education**. This workshop is annually organized by the Center for Language and Speech Processing, Johns Hopkins University, Baltimore, USA headed by Frederick Jelinek, the founder and world leader in statistical methods in speech recognition and language modelling. Project was running under the leadership of Jan Hajic and some of the most recognized scholars in the field, such as Eric Brill, Michael Collins and Lance Ramshaw, were team senior members; I was a team member as a graduate student. I closely cooperated with Eric Brill (now at Microsoft, Redmond, USA) on so-called 'superparser' to explore methods for combining parsers (automatic procedures for syntactic analysis). My project proposal on building a 'supertagger' was rewarded by the **Post Workshop'98 Research Project Award**, which allowed me to study new horizons of tagging.

I defended my PhD thesis on Czech language tagging in 2000 (thesis readers: prof. Frederick Jelinek, Johns Hopkins University, Baltimore, USA, prof. Petr Sgall, Charles University in Prague) and the first version of the Prague Dependency Treebank was published by the Linguistic Data Consortium (LDC), Philadelphia, Pennsylvania, USA in 2001. The experience I got while working on my PhD thesis, while building the PDT and during the Summer Workshop'98, gave me a broader insight into morphology and syntax with regards to both data and tools.

Since 2004, I am **the PI of a five-year project** "Resources and Tools for Information Systems" where I supervise conversion of the internal format and annotation schemes of the Czech Academic Corpus in a way that they are compatible with PDT. Currently, the second version of the Czech Academic Corpus together with the tools handling morphology and syntax is published by LDC. It facilitates the possibility of integrating CAC directly into PDT and thus to get more training data for tagging and parsing. There is no

other language “having at its disposal” the annotated texts of such carefully formulated annotation guidelines and a significant volume – all together, PDT and CAC consist of 2.6 million words.

Having a deep experience with building the annotated corpora, I am interested in the idea of using them outside their original context. The idea to build an **electronic exercise book based on PDT** to learn and practice Czech morphology and syntax became a topic of the Master thesis that I supervised and that was successfully defended in 2005. The idea was recognized as a novel one, as no other annotated corpus has been used for such purpose.

During the building the annotated corpora and using them as training data for machine learning approaches, I have learned that corpus annotation is needed even though it is an expensive activity. Therefore, I am currently interested in an alternative way of **annotation through the on-line games**. I am also studying the linguistic phenomena, which cross the sentence boundary and that contain information concerning the contents of the document. I am particularly interested in degrees of salience of items in the stock of shared knowledge the speaker assumes (s)he shares with the hearer. Text summarization, text classification, information retrieval systems and question answering systems are examples of the applications that can benefit from this research issue.

In the academic years 1998/99 and 1999/2000, I was running an undergraduate and a graduate course on **Statistical Methods in Natural Language Processing** at the Faculty of Mathematics and Physics, CUNI. Since 2003 till now, I am running an undergraduate and a graduate course on **Introduction to Machine Learning in Natural Language Processing** at the same faculty. These courses are (were) attended by 10 students in four hours a week. The students are supposed to work on a project related to a selected task of natural language processing and one of the students, Jana Kravalova, was awarded for her project at the University Student Competition in Mathematics and Computer Science, 2008. The course on Machine Learning is open in English, because it is attended also by the international students coming from the European Masters Program in Language and Communication Technologies.

In a period 2005-2007, I **supervised the Master** thesis of Ondrej Kucera on building a corpus-based exercise book of Czech morphology and syntax. His work was awarded twice: (i) the Third Best Demo at the Demonstration Session at the 2005 Conference on Empirical Methods on Natural Language Processing – Human Language Technologies, Vancouver, British Columbia, Canada and (ii) a finalist of the 4th Student Research Competition in Informatics and Information Technologies organized by the Czech chapter of ACM, 2006. Ondrej is still working on the exercise book as a PhD student under my supervision.

In the years 2000-2007, I **supervised two PhD students** – Otakar Smrz (defended in 2007, now at IBM, Prague) and Zdenek Zabokrtsky (defended in 2005, now at the Institute of Formal and Applied Linguistics). Otakar’s PhD thesis is on the functional Arabic morphology and Zdenek’s PhD thesis on the valency lexicon of Czech verbs.

I regularly serve as a **paper reviewer** for major Computational Linguistics conferences and I also contribute to the organization of international scholarly meetings. My largest task was to **organize** an international two-day tutorial “Prague Treebanking for Everyone” to introduce a family of the Prague treebanks to students and researchers. Tutorial was attended by hundred participants.

I believe – as attested above – that I have the skills, experience and leadership ability to successfully lead the proposed research under the ERC Starting Grant and to achieve significant results, going beyond the current knowledge of my research field.

Awards

Josef Hlavka’s Award for the best students of Charles University in Prague	1996
Post Workshop’98 Research Award, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, USA	1998