Research statement

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 participated 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 my research on applying corpus-based methods to tagging under supervision of Jan Hajič. All the largely used **tagging** procedures were originally designed for English. These procedures are corpus-based (i.e. supervised) 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.[1] 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 syntax to semantics. 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

1 http://ufal.mff.cuni.cz/pdt/

for the project in the prestigious **NSF Workshop** annually organized by the Center for Language and Speech Processing, Johns Hopkins University, Baltimore, USA: **Language Engineering for Students and Professionals Integrating Research and Education**.[2] The 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 eBay Research Labs, 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.

I was the **PI** of a five-year project (2004-2008) "**Resources and Tools for Information Systems**" where I supervised conversion of the internal format and annotation schemes of the Czech Academic Corpus in a way that they are compatible with PDT. Finally, the second version of the Czech Academic Corpus together with the tools handling morphology and syntax has been published by LDC.[3] 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.[4] The idea was recognized as a novel one, as no other annotated corpus has been used for such purpose.

---

2 http://www.clsp.jhu.edu/workshops/archive/ws98-summer-workshop/
3 http://ufal.mff.cuni.cz/cac
4 http://ufal.mff.cuni.cz/styx

During the building the annotated corpora and using them as training data for machine learning approaches, I have learned that corpus annotation, i.e. building treebanks is needed even though it is an expensive and very time consuming activity, thus we have been seeking for an alternative cheaper and faster way, like crowdsourcing. Therefore, in 2009-2010 I was interested in an **alternative way of annotation** through the **on-line games**. Namely, we built the game portal LGame: Play the Language consisting of three linguistically motivated on-line games.[5]

Currently, I'm involved in three research projects:

- **INTLIB**[6] – The aim of this project is to provide a more efficient and user-friendly tool for querying textual documents than full-text. We work with a collection of documents related to a particular problem domain, mainly legislative and environmental. In the first phase we extract from the documents a knowledge base -- a set of entities and their relationship -- using natural language processing tools. In the second phase we represent extracted data according to the Linked data principles. The project presents a joint work of MFF UK and Sysnet, Ltd. Both my PhD students, Vincent Kríž and Ivana Lukšová are involved. The jTagger[7] and RExtractor[8] systems have been designed and implemented to detect entities and their relationships.

- **Čapek**[9] – One way of teaching grammar, namely morphology and syntax, is to visualize sentences as diagrams capturing relationships between words. Similarly, such relationships are captured in a more complex way in treebanks serving as key building stones in modern natural language processing. The purpose of our work is to explore possibility to get sentence diagrams produced by students and teachers. Mainly, we (i) design a tool for drawing sentence diagrams, so call the Čapek editor, that attract teachers to use it in language classes and encourage students to use it for practicing on their own; (ii) ensure that the quality and quantity of the obtained data satisfy requirements for applying supervised learning methods; (iii) design transformation rules to enlarge a volume of training data for machine learning. In our pilot study, the object language is Czech, where sentence diagrams are part of elementary school curriculum. I'm addresing this task together with my colleague Jirka Hana and my PhD student Ivana Lukšová. The project is partially funded by the Charles University Grant Agency, Ivana's grant *An alternative way of getting more annotated linguistic data,* no. 1568314.

5 http://ufal.mff.cuni.cz/tools/lgame
6 http://ufal.mff.cuni.cz/intlib
7 http://ufal.mff.cuni.cz/jtagger/
8 http://odcs.xrg.cz/devel-rextractor
9 http://ufal.mff.cuni.cz/capek

- **Salience analysis** – I am studying the linguistic phenomena, which cross the sentence boundary and that contain information concerning the contents of the document. I am particularly interested 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** together with my colleague Martin Holub at the same faculty. This one-semester introductory course provides theoretical background of and key algorithms from the field of machine learning explained on multidisciplinary applications. The lab sessions are application-dependent and they accompany the lectures The aim of the labs is an acquisition of practical experience from application of machine learning approaches on problems from the field of natural language processing. The course 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 August 2013, I and Martin Holub were giving an introductory **course** *A Gentle Introduction to Machine Learning in Natural Language Processing using R* at the 25th European Summer School in Logic, Language and Information (**ESSLLI 2013**) organized by Heinrich Heine University in Düsseldorf, Germany.

In a period 2005-2007, I supervised the Master thesis of Ondřej Kučera 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.

In the years 2000-2007, I supervised two PhD students – Otakar Smrž (defended in 2007, now at Seznam.cz, Prague) and Zdeněk Žabokrtský (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. Currently, I supervise two PhD students, namely Vincent Kríž and Ivana Lukšová. Both of them are approaching the task of entity and relation extraction from unstructured documents of various domains.

I am supervising many undergraduate projects which introduce the basic and more advanced concepts of natural language processing to the students and which apply the data, tools and ideas developed so far to new progressive topics.

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” in Dec 2006 to introduce a family of the Prague treebanks to students and researchers. Tutorial was attended by hundred participants.

I have done research in natural language processing so far; however my skills and experience with data processing and machine learning approaches can be applied to any other field, if applicable.