BRNO UNIVERSITY OF TECHNOLOGY

FACULTY OF INFORMATION TECHNOLOGY

Grammars for Natural Languages: Methods, Tools and Lexical Resources for Their Efficient Processing

HABILITATION THESIS

2005 **Pavel Smrž**

Abstract

This habilitation thesis deals with the research on grammars for natural languages. It comprises my articles and papers on grammars for Czech and other languages, approaches to their integration into efficient parsers and lexical resources and relevant tools supporting the analysis. The first part deals with the concept of metagrammar — a special kind of grammar that focuses on modularity and reduction of rules needed to describe the grammar of free word order languages. The methods of efficient parsing with the metagrammar as well as best-analysis selection techniques are discussed next. The applications of the developed parser are also tackled. The second part deals with morphological and lexical resources and tools for their building. It covers Czech morphological analyzer, wordnet and other lexical databases, ontologies, and word sketches. Various tools for creation of these resources (manual as well as fully automatic) are also presented.

Keywords

metagrammar, parsing, language resources, ontologies

Acknowledgements

The work described in this thesis could not be done without the help of my colleagues from the Natural Language Processing Laboratory, Faculty of Informatics, Masaryk University, Brno. I would like especially thank to Karel Pala, Pavel Rychlý, Aleš Horák, Vladimír Kadlec, Martin Povolný and Radek Sedláček. I am also obliged to Anna Sinopalnikova for her support and cooperation. Last but not least, I am really obliged to Jan Černocký, Pavel Zemčík and other colleagues from the Faculty of Information Technology, Brno University of Technology, who had lent me their hand when I needed it. Many thanks to my parents and all my family for their love.

Contents

Chapter 1

Introduction

1.1 Motivation

Natural language is a system of signs combined according to rules of a grammar for the purpose of communication. The term “grammar” is used for both — an inherent (implicit) framework of sign-combining capabilities of a language as well as an explicit specification of the system aiming at the (grammatical) description of the language. The latter sense is taken into account in the following sections.

If a grammar is to be employed in an automatic NLP (natural language processing) system, one needs to look for a formal grammar, i. e. a formal specification of the language rules that can be transformed into a parser. There are two techniques to come by a formal grammar. The first approach to the grammar construction is to build it manually. Linguists specify the grammar rules and test them on a given set of data. Large-coverage manually-defined grammars for English and a few other languages are available today and are successfully built into wide-ranging systems.

Alongside the growing popularity of machine-learning methods in NLP in the last two decades, another approach to grammar creation appeared. It consists in an automatic induction of the grammar rules from available training data [KM02]. One can also resign completely from standard grammars and focus rather on the parsing task itself. The data-oriented parsing [Bod99] follows this line. Nowadays, useful parsers are based on the data automatically extracted from large, syntactically annotated corpora (treebanks) such as Penn treebank [MSM94], and are regularly applied for various tasks, e. g. for language modeling [CJ98].

The system presented in this thesis combines the manual definition of a metagrammar structure (see Chapter 2) with the learning of grammar-rule probabilities from training data. This organization preserves the direct interpretability of the grammar (and the possibility to “tune” it manually when there is not enough data for specific language phenomena) and offers an appropriate framework for the development of parallel grammars. On the other hand, this configuration enables construction of a stochastic grammar [Boo69] that can take advantage of vast syntactically-tagged data available.

Czech is one of a few languages that can benefit from a large treebank. PDT (Prague Dependency Treebank) [Haj98] is a comprehensive resource covering several levels of language analysis — from simple morphology to the tectogrammatical level [SPH04]. Machine learning methods have been applied to parsing based on the data extracted from PDT in last years (see, e. g., [CHB$^+$99]). However, the later experiments showed that the standard stochastic parsing methods cannot be easily improved by deep linguistic knowledge (e. g. by verb valency lists [ŽL04]) which seems to be necessary for an acceptable parser performance. Apart from other goals, the research summarized in this thesis aims at a better understanding of the integration of language-specific features to parsing. I strongly believe that a substantial part of the grammar, which is specified manually at present, could be extracted from data automatically in future and that linguistically-oriented approaches and machine-learning techniques (traditionally taken as antagonistic in NLP [Chu04]) can be combined and will be finally brought into accord in a future NLP system.

To build a formal grammar, one needs a formalism that specifies the way of expressing rules and other parts of the grammar. This specification usually implies a more or less complete theory of language (or, at least, of its syntactic component). There are numerous grammar theories and related formalisms developed in last decades that can be employed. The most popular ones are LFG — Lexical Functional Grammars [Bre01], HPSG — Head-driven Phrase Structure Grammars [PS94], and LTAG — Lexicalized Tree-Adjoining Grammars [Jos87, SAJ88]. Within their frameworks, broad-coverage grammars were defined for English [XTAG01, LGO04, Fli00] and other languages [MK00, Sie00, BDK$^+$02]. Various analyzers based on these grammars were developed to parse texts in the respective languages too [MK93].

Faced with the task to develop a grammar for Czech, we should choose the best formalism for our purposes. If the objective is to directly apply the grammar in a form of an efficient and robust parser (see the discussion in Chapters 4–7), "the best" could mean a formalism which offers the most appropriate mechanisms to cope with the language-specific features. Unfortunately, none of the discussed frameworks has been used for a large project in Czech till now. To the best of our knowledge, the same holds for other Slavic languages as well as for similar languages (with free word order and rich morphology) belonging to other language families. Only limited parts of the grammar (or just "toy" grammars) are available for languages similar to Czech (see, e. g., the survey of HPSG grammars for Slavic languages [Prz00]). No large-scale grammars tackling all major language phenomena are available for those languages. Moreover, the most popular formalisms were designed to reflect the original theory of language and to define the grammar in a declarative way rather than aiming at efficient parsing. Special techniques increasing the efficiency of the relevant parsers were developed in last years only [Usz02] and the mechanisms dealing with free word order still wait for their implementation or testing (cf. the situation of the HPSG parser TRALE described in [Ric04]). All these findings influenced the research on the metagrammar covered in this thesis.

The theoretical background to capture the word order in Czech (and other languages) is provided by FGD — Functional Generative Description [SHP86]. The Prague Linguistic School has a long tradition of research on communicative dynamism and the study of connection between the word order and the topic-focus structure of sentences [HSS94]. FGD states a theory rather than a particular grammar formalism. A recent work of our colleagues explored the possibility to employ the formalism of categorical combinatorial grammars [Kru01]. Even though our system does not follow the same approach, it shares many features and reflects the same aspects of FGD. The strict orientation to dependency-based analysis influenced the representation in the form of dependency graph (a combination of several non-conflicting dependency trees) used by the implemented tool. The tectogrammatical-level analysis inspired the mechanism of so called intersegments that can play the role of elided items in the analyzed sentences. Also, the described head-driven parsing algorithm takes advantage of the explicitly given dependency relations. However, the dependency orientation of the metagrammar is "unorthodox" in the sense it concedes constituency in the analysis of some language phenomena. For example, coordinative groups are analyzed as constituents that are taken as special elements in the higher levels of the analysis. It enables natural specification of agreement rules, e. g. the constraint on the relevant predicate in plural.

The metagrammar presented in this thesis offers a general framework that focuses on modularity and reduction of rules needed to describe the grammar of free word order languages. It combines constituency and dependency approaches. The governing node (head) is specified in each metagrammar rule and the output of the analysis is presented in the form of a dependency tree (or a packed graph in the case of ambiguity). The context-free backbone of the grammar is linked to contextual constraints defined on the base of metarules. Such a design allows very efficient parsing which plays a crucial role in the employed application-oriented approach. It also enables the metagrammar to serve as a source of linguistic description of many phenomena that are not currently covered for Czech formally. This way, a subset of the metagrammar rules was used to generate a base for an HPSG grammar fragment in a previous student project [Nov03]. I am glad to make the results of my research available to the colleagues working in this area to consult them in the creation of new natural language grammars.

From the very beginning, all my work on the metagrammar was performed with the vision of building a practical application — an efficient large-coverage parsing system. Among others, this goal implied the integration of an efficient morphological analyzer for languages with rich morphology. As there was no appropriate morphological tool available for Czech at that time, we exerted on this task and devoted a considerable amount of time to morphological tools and resources that helped to develop them and to test their functionality. We created the first morphologically-annotated corpus for Czech — DESAM (joint work with K. Pala, P. Rychlý and other colleagues from NLPlab FI MU). The work on manual disambiguation of the morphological tags showed that many morpho-syntactical phenomena in Czech deserved special attention and that many problems were still to be solved. We also recognized that the tools available at that time and used in our work did not satisfy our needs and could not easily become a part of a future syntactic analyzer. Thus, the next step was to focus on the morphological analyzer. We developed a new tool `ajka` (joint work with R. Sedláček) which has soon become one of the most frequently used applications in many other projects. `ajka`, with its recent extension to derivational morphology [Sed05], is still unique in many respects and provides excellent results in terms of efficiency and coverage.

In addition to the rich morphology, the other key feature of Czech and other Slavic languages, the syntactic analyzer must cope with, is the word order. The general form of the metagrammar is influenced by the ID/LP (Immediate Dominance/Linear Precedence) formalism [GP81]. Instead of immediate dominance, we focus on the dependency relation. The governing elements are identified in the metagrammar rules and this information is used to define the dependency shared forest resulting from the analysis. On the other hand, special constructs of the metagrammar enable defining linear precedence. The progressive development of the Czech metagrammar clearly showed that this specification can be seen as a strict constraint only for a very limited number of language phenomena. For example, it can be employed in the rules for enclitics but not in that of noun modifiers. Therefore, linear-precedence constructs in the metagrammar are interpreted as preferences and the actual corpus training data govern their use.

The flexible metagrammar word-order rules and the quest for a maximal coverage of the grammar necessarily influence the results of the analysis. We also aim at a robust parser [JvN01] able to return the best possible partial analysis even for incorrect input sentences. All these factors impact the enormous number of (potential) analyses that can be produced by the parser. It is not a trivial task to achieve an efficient analysis for such a system. The efficiency issue is stressed here as it forms the golden thread of all the development work described in this thesis. The current head-driven chart parser is a result of long-term optimization that led to a very efficient system. The comparison with other available tools on the standard evaluation set of very large grammars proved that our parser is one of the fastest systems all over the world.

As it was already stated, the primary way to deal with many resulting analyses is to employ statistics (estimated on frequencies extracted from PDT). What needs to be specified is what level the statistics should operate on. The most straightforward answer is to compute frequencies based on the pre-defined grammar categories. This approach works well for the most frequent language phenomena (e. g., to determine the general preference for genitive relations). However, other phenomena call for more detailed information of word-combination potential and for lexicalization of particular grammar rules. Full lexicalization is problematic: the Zipf law [Zip35] says that the data sparseness problems are inevitable and that no corpus can be large enough to provide data for the estimation of all the probabilities necessary for the best analysis selection. Inspired by the class-based approaches used in the language modeling [Jel90], we decided to overcome the data sparseness in the probability estimation by grouping words to their natural classes. These considerations determined a completely new direction of our research in the area of language resources. Various dictionaries, thesauri and valency lexicons have been created. They were built either manually using the developed editing tools, or with the help of automatic procedures that extract data from large text corpora. Both approaches are tackled in this thesis:

A considerable effort was put into the development of the Czech wordnet and tools for its building (joint research with K. Pala, P. Rychlý and A. Horák). The core of the lexical database was developed in the EuroWordNet project [Vos98] and it was substantially extended and improved within the BalkaNet project [TCS04]. Nowadays, the Czech wordnet covers not only the standard semantic relations such as synonymy, antonymy, hypernymy, meronymy, etc. but special mechanisms are also provided for language-specific relations — diminutives, regular derivations between verbs and corresponding nouns and many others. It is one of the largest wordnet databases arising from the English original.

We closely cooperated with other developers of national wordnets, interchanged knowledge and experience. Especially fruitful was the joint work with Russian colleagues from the St. Petersburg State University. We assisted and supported the development of RussNet [AMS$^+$02]. This resource is unique in many respects; it comprises several features not seen in the other wordnet-like databases. It is based completely on empirical data. Corpus-based methods have been applied in its creation. Due to the similarity between Russian and Czech, the RussNet project benefited significantly from the experience, resources and tools gained for the Czech language [SS04].

Both the mentioned large wordnet-oriented projects — EuroWordNet and BalkaNet embodied a parallel building of several national wordnets. This task is usually performed by groups of enthusiasts that devote their entire energy to cover as much of the semantic relations as possible. However, the resulting lexical databases do not often reflect this enormous effort to the full extent. Having the background in software engineering, I therefore paid attention to the engineering side of the wordnet development and devised a set of quality-assurance procedures for these purposes. They were successfully applied in the final stage of the BalkaNet project and contributed to the high quality of all the wordnet databases developed within it.

A key issue in the development of lexical resources is the availability of supporting tools for their building. As there was no appropriate editor at the beginning of the BalkaNet project (see the discussion of this situation in Chapters 10 and 11), we had to focus on this task too (joint work with A. Horák, T. Pavelek and P. Rychlý). Visdic — a wordnet editor and browser — achieved a great success. Originally designed as merely a supporting tool for the BalkaNet project, it soon became a standard tool for many projects aiming at language resource development. Among others, it is currently employed in the development of RussNet, extension works on the Romanian wordnet, the creation of a core of the Estonian one etc.

Despite its growing popularity, Visdic deployment also revealed limitations of the tool. The extended relations going far beyond the English wordnet brought new requirements that were not envisioned in the earlier design phases. It became obvious that the design of Visdic is not flexible enough to incorporate all the necessary new functionality. Rather than re-implement the old tool we decided to develop a completely new one that would correspond to the current trends and standards in lexical database representation (joint work with M. Povolný). DEB (Dictionary Editor and Browser) is based on the multi-layer architecture. It takes advantage of advanced technologies supporting the XML database format and its transformations. The applicability is not limited to wordnet databases. The flexible presentation level enables specifications for any kind of relations among concepts designated by the language expressions. At the same time, special mechanisms manipulating the XML data provide efficient search and retrieval for large lexical databases.

The wordnet-like lexical resources can be seen as the simplest form of ontologies. In computer science, this term stands for the formal and machine-readable representation of concepts and relations among them. Ontologies can be used for various tasks in natural language processing and, if detailed enough, they can be directly applied in language analysis. The work on the advanced ontology tools brought us also to the area of the Semantic Web where ontologies play the major role. The Berners-Lee's vision of the data on the web understandable by machines [BLHL01] and its subsequent elaborations and definitions of related standards, e. g. OWL — Ontology Web Language — provided a new domain of application for our research. It was centered on the relation among ontologies, verb frames, and semantic roles. It has been shown that ontologies, even in their current imperfect form, offer a base for lexico-semantic preferences and that the semantic roles participating in valency frames can be anchored in particular ontological concepts [PS04]. Moreover, the research proved that the previous applications of the parser for the analysis of verb valencies and subsequent translation of the analyzed sentences to constructions of Transparent Intensional Logic (TIL) [Tic94] can be supported by the TIL types specified in the valency frames.

Valency frames form a connecting link between manual and fully automatic building of lexical databases. The automatic extraction of semantic relations is usually based on the distributional hypothesis, which states that words with similar meanings tend to occur in similar contexts. This idea is also inherently contained in another result of my research (joint work with P. Rychlý) — word sketches for Czech. Word sketches are short corpus-based summaries of a word's grammatical and collocational behavior. Their creation consists in a massive extraction of semantically-related words from very large corpora. The method is based on pre-defined grammatical patterns (extracted from the metagrammar in our case). Our research proved that this valuable resource can be created not only for English, but also for languages with free word order. It has been also shown that the available tools allow employing the same procedure for the development of Russian word sketches.

As described in the following sections, I worked not only on the metagrammar and the parser for Czech, but participated in the development of various lexical resources and tools for their creation. It is obvious that the discussed tools and resources have broader application potential than just supporting a language parser. Many of them were applied in several projects that are not directly connected to the area of syntactic analysis or take the syntax just as one of many integrated components. One of these projects I am currently involved in is the application of the developed tools and resources in e-learning. An application platform that integrates the results discussed in this thesis was introduced in [Smr04a]. It shows that natural language processing technologies can bring a new quality to the computer-mediated teaching and learning. It is also apparent that more natural way of communication brought by NLP-aware learning management systems can boost the popularity of e-learning. This direction presents a field of my current research [Smr04b, SF05].

1.2 Overview

This subsection contains an overview of the following chapters. Some topics are not covered in detail. In these cases, there are references to my papers that can provide background information for the respective topics. Note also that the work on several discussed problems is collective. My participation is exactly specified in part 4 of the application for habilitation.

The next section contains my article *A Parallel Metagrammar for Closely Related Languages — A Case Study of Czech and Russian* that is to be published in the Research on Language and Computation Journal, the special issue on Shared Representation in Multilingual Grammar Engineering. The work offers an answer to the question to what extent the representation used for Czech could be applied to Russian and which levels of representation should be shared across the grammars. It demonstrates the advantages of the developed formalism for parallel coverage of Czech and Russian. The generalized metagrammar constructs enable sharing substantial parts of the grammar description between the two closely related languages.

Section 3 brings the paper *Grammatical Heads Optimized for Parsing and Their Comparison with Linguistic Intuition* (V. Kadlec and P. Smrž, presented at TSD 2004). It emphasizes the importance of machine-learning techniques for the automatic parsing of natural languages. Dependency grammars identify governing elements for the relation of dependency (parent nodes in dependency trees). Grammars for head-driven parsing algorithms specify the elements — heads of the rules — from which the analysis should start. The standard procedure assumes that the governing element also serves as the head for parsing. However, our work shows that the positions of heads driven by the linguistic intuition are not always optimal for parsing. If the efficiency is in the focus, the head positions should be learned from training data. Some examples of the differences between "linguistic" and "empirical" heads are given in the paper.

The paper *Probabilistic Head-Driven Chart Parsing of Czech Sentences* (coauthored by A. Horák, presented at TSD 2000, Section 4 in this thesis) also links the grammar development, focusing on the declarative linguistic description, to the development of an efficient parser for Czech. The former is demonstrated by the advanced metagrammar constructs introduced in the paper, the later by the implemented head-driven chart parser able to analyze long sentences (more than 40 words) in very short time. This paper also shows the connection between the metagrammar created manually and the machine learning approach to the selection of the best analysis. As the number of potential syntactic trees for some sentences can be very high, the parser was supplemented with statistics learned from PDT. The statistics are involved in the process of sorting out the edges from agenda in the order that leads directly to *n* most probable analyses.

Two following papers — *Efficient Sentence Parsing with Language Specific Features: A Case Study of Czech* (A. Horák and P. Smrž, presented at IWPT 2001, Section 5 in this thesis) and Best Analysis Selection in Inflectional Languages [HS02] (presented at COLING 2002, not included in this thesis) — explore the optimization techniques employing detailed information about language-specific properties. The first method involves a special type of merging of the computed values in the standard chart-parsing algorithm. The reduction of memory and time requirements is possible due to the agreement conditions in Czech (e. g., the agreement among the governing noun and its adjective attributes). Other language-specific features are integrated in the best analysis selection. "The best" output is judged by a probabilistic figure of merit. The term "figure of merit" is usually used to refer to a function that prunes implausible partial analyses during parsing. As the parser enables extremely fast computation of all possible parses, the figure of merit is rather taken as a measure bounding the true probabilities of the complete parses. Particular language features reflected in the implemented figures of merit include the frequency of syntactic constructs represented by the rule probabilities pre-computed from PDT and an augmented *n*-gram model based on the occurrence of adjacent lexical heads standing for the corresponding subtrees. Another crucial pieces of knowledge integrated in the work are the verb valency frames used for probabilistic ordering of parsing output. The list of valencies and their semiautomatic acquisition are tackled later in the thesis in the context of building of transparent-intensional-logic representation of analyzed sentences.

Two possible modifications of the standard form of head-driven chart parser are discussed in the paper *How Many Dots Are Really Needed for Head-Driven Chart Parsing?* (coauthored by V. Kadlec, accepted for publication in the Proceedings of Sofsem 2006, Section 6 in this thesis). Both techniques reduce the number of chart edges by modifying the form of items (dotted rules). The presentation takes advantage of parsing schemata — formal algebraic structures appropriate for description of parsing algorithms. Testing on a standard set of large grammars proves that the number of edges in the resulting chart can be significantly decreased.

The paper *Incremental Parser for Czech* (coauthored by V. Kadlec, published in the Proceedings of the 4th International Symposium on Information and Communication Technologies, Section 7 in this thesis) bridges the topics of the development of the parser and its applications. Several NLP tasks require incremental creation of the analysis. For example, structured language models that can be employed in automatic speech recognition or predictive writing of messages on cellular phones usually expect the integration of an incremental parser. The paper demonstrates how the pruning of contextual constraints specified in the metagrammar can be efficiently evaluated in the incremental mode. The algorithm takes advantage of the limited number of values that can be produced for each type of the constraint. Instead of pruning the original packed share forest, the parser builds a new forest of values resulting from the evaluation.

The use of the developed metagrammar and the parser for the analysis of verb valencies and subsequent translation of the analyzed sentences to constructions of Transparent Intensional Logic (TIL) is discussed in the paper *Determining Type of TIL Construction with Verb Valency Analyser* (coauthored by A. Horák, presented at Sofsem 1998, Section 8 in this thesis). The existing list of verb valency patterns is validated and extended by means of partial syntactic analysis (using the previous implementation of the parser, see also P. Smrž and A. Horák — Partial Syntactic Analysis as a Tool for Semiautomatic Verb Valencies Acquisition and Checking [SH98], TSD 1998, not included in this thesis). The next step is to assign logical (TIL) constructions that correspond to the verb meaning. This process generates input for the normal translation algorithm [Hor01], a base for the logical semantic analysis of the Czech language based on TIL.

Another "standard" application of the parser for Czech lies in morphological disambiguation of texts. Even a parser with the precision of about 80 % can provide accurate results for morphological disambiguation. The analysis errors often go for the structural level (e. g., prepositional attachment problems) and do not influence the morphological one. The strengths and weaknesses of the disambiguation tools based on grammars are discussed in the paper New Tools for Disambiguation of Czech Texts [SŽ98] (coauthored by E. Žáčková, not included in this thesis). The most important supportive tool that had been developed was the morphological analyzer `ajka`. The paper *A New Czech Morphological Analyser* `ajka` (R. Sedláček and P. Smrž, presented at TSD 2001, Section 9 in this thesis) introduces the ideas behind its design and discusses some implementation issues. Ajka was integrated with the developed syntactic parser and weighted in the large coverage of the tool.

The implemented Czech parser selects the most probable analysis based on frequencies obtained from training data. This computation can be fully lexicalized, i. e., the frequencies of grammatical relations can be determined for individual words. Unfortunately, the available training corpora do not suffice for the full lexicalization — one needs to cope with the data sparsity problem. Inspired by the use of English WordNet [MBF$^+$90] for such purposes, we took advantage of two European projects — EuroWordNet and BalkaNet and developed the wordnet database for Czech. The paper *Building Czech Wordnet* (K. Pala and P. Smrž, published in the Romanian Journal of Information

Science and Technology, Section 10 in this thesis) reports on this work in the context of the BalkaNet project. The resources and tools employed in various phases of the development are introduced and the resulting database is described. The specific features of Czech as a language with rich inflectional and derivational morphology and their impact on the wordnet database structure are also studied.

To avoid the problems connected with the engineering phase of the lexical-database creation, we also focused on the quality assurance procedures. A part of the work on language resources concentrated on software tools that can help to manage the coordination of the teams collaborating in the development of multilingual lexical database. This aspect is studied in my paper *Lexical Databases as a Base for Broad Coverage Ontologies* (presented at the AAAI Workshop on Ontologies and the Semantic Web, AAAI 2002, Section 11 in this thesis). It shows how the lexical knowledge bases can be shared and combined, how to merge independently developed parts of ontologies, and how to check inconsistencies and report errors.

The quality management is further explored in two my papers Quality Control and Checking for Wordnet Development [Smr04c] (published in the Proceedings of the Second Global Wordnet Conference, not included in this thesis) and *Quality Control and Checking for Wordnet Development: A Case Study of BalkaNet* (published in the Romanian Journal of Information Science and Technology, Section 12 in this thesis). The designed and implemented quality-assurance procedures were successfully applied in the context of the development of all the five languages of the BalkaNet project.

In the BalkaNet project, a new format was designed for the wordnet and other semantic databases that became a de facto standard in the area of wordnet development. As there were no tools able to work efficiently with the databases and to satisfy other requirements (e. g. multiplatform system — MS Windows as well as Linux) the Czech team engaged in the design and the implementation of the editor and browser for general lexical databases in the XML format. DEB — a unique tool allowing fast development of lexical databases was introduced in the series of papers Lexical Databases in XML: A Case Study of Up-Translation of the Dictionary of Literary Czech Language [Smr02] (presented at EURALEX 2002, not included in this thesis), DEB — A Dictionary Editor and Browser [SP03] (coauthored by M. Povolný, presented at the 4th Papillon Workshop, not included in this thesis) and *Lexical Databases in XML* (coauthored by M. Povolný, presented at the EACL Workshop on Language Technology and the Semantic Web — NLPXML-2003, Section 13 in this thesis). DEB (Dictionary Editor and Browser) is based on the client-server architecture and gets over Visdic in the support of all the range of XML technologies and related W3C standards (XSLT, XML Schema, XPath, DOM, etc.). A special attention is paid to a feature which brings the efficiency of retrieval. It extends a standard XSLT processor with the ability to obtain additional data from the dictionary server through the mechanism of nested queries. It also shows that the emphasis on the standards facilitates the connection to other linguistic tools such as corpus managers, morphological analyzers and parsers (see also P. Smrž, M. Povolný, A. Sinopalnikova: OASIS — A New Tool for the Transformation of XML Knowledge Resources into OWL, ISWC 2004 [SPS04], not included in this thesis).

The developed metagrammar is primarily intended for the deep syntactic analysis of Czech. Nonetheless, it also provided a basis for the grammatical patterns used in the definition of wordsketches for Czech. The paper The Wordsketch Engine [KRPP04] (A. Kilgarriff, P. Rychlý, P. Smrž and D. Tugwell, presented at EURALEX 2004, not included in this thesis) discusses the advantages of this new tool that can facilitate lexicographic work and demonstrates how it can be applied in languages with free word order. Particular attention has been paid to the agreement constraints that are typical for Czech. Our metagrammar constructs can be easily transformed to the special type of the grammar patterns shown in the paper. The definition of the grammatical patterns is tackled in the paper *Manatee, Bonito and Word Sketches for Czech* (P. Rychlý and P. Smrž, presented at the Second International Conference on Corpus Linguistics in Saint-Petersburg, Section 14 in this thesis). It also suggests the next step of our work — building the wordsketches for Russian.

The previous paragraphs summarized my research in the field of natural language processing. The developed metagrammar proved to be valuable in practical applications. Also, the described parsing techniques provide excellent results on large ambiguous grammars typical for Slavic languages. It has been shown that various lexical resources can be integrated into the parsing process and that they can help in selecting the best (most probable) analysis if there are not enough training data to apply standard stochastic approaches. The designed and implemented tools facilitate manual as well as semi-automatic creation of these tools and enable their employment.

Natural language understanding by machines is an extremely ambitious task. My research in the previous years has contributed to the promotion of the automatic processing of Czech and other languages. It is obvious that a lot of work remains to be done. The new topic of my current work is the structured modeling for speech recognition and “robustification” of the current syntactic analyzer to cover spontaneous speech. I believe that the previous effort on the grammars, language resources and various tools for their building that are covered in this thesis will help to push forward the frontiers in this area.

References to My Papers

[HS02] A. Horák and P Smrž. Best analysis selection in inflectional languages. In *Proceedings of the Nineteenth International Conference on Computational Linguistics (COLING)*, pages 363–368, Taipei, Taiwan, 2002. The Association for Computational Linguistics.

[HS04] A. Horák and P. Smrž. Visdic — Wordnet browsing and editing tool. In *Proceedings of the Second International WordNet Conference (GWC)*, pages 136–141, 2004. ISBN 80-210-3302-9.

[KRPP04] A. Kilgarriff, P. Rychlý, P. Smrž, and D. Tugwell. The sketch engine. In *Proceedings of the Eleventh EURALEX International Congress*, pages 105–116, Lorient, France: Universite de Bretagne-Sud, 2004. ISBN 2952245703.

[PS04] K. Pala and P. Smrž. Top ontology as a tool for semantic role tagging. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, pages 1897–1900, Lisbon, Portugal, 2004. Centro Cultural de Belem. ISBN 2-9517408-1-6.

[SF05] P. Smrž and M. Fapšo. Searching in the recordings of lecture. In *Proceedings of the Workshop on Technologies in eEducation*, pages 21–26, Prague, 2005. (in Czech).

[SH98] P. Smrž and A. Horák. Partial syntactic analysis as a tool for semiautomatic verb valencies acquisition and checking. In *Proceedings of the First International Conference on Text, Speech, Dialogue (TSD)*, pages 123–128, 1998.

[Smr02] P. Smrž. Lexical databases in XML: A case study of up-translation of the Dictionary of Literary Czech Language. In *Proceedings of the Tenth EURALEX International Congress*, pages 729–734. Copenhagen, Denmark: CST, Copenhagen, Denmark, 2002. ISBN 87-90708-09-1.

[Smr04a] P. Smrž. Integrating Natural Language Processing into E-learning — A Case of Czech In *Proceedings of the Workshop on eLearning for Computational Linguistics and Computational Linguistics for eLearning, COLING*, pages 1–10. Geneva, Switzerland: Universite de Geneve, 2004.

[Smr04b] P. Smrž. Integrating ontologies into learning management systems — A case of Czech. In *Proceedings of the Workshop on Ontologies, Semantics, and E-learning (WOSE), On the Move to Meaningful Internet Systems, OTM 2004 Workshops*, pages 768–772. Berlin, Germany: Springer, 2004. ISBN 3-540-23664-3.

[Smr04c] P. Smrž. Quality control for wordnet development. In *Proceedings of the Second International WordNet Conference (GWC)*, pages 206–212, 2004. ISBN 80-210-3302-9.

[SP03] P. Smrž and M. Povolný. Deb — a dictionary editor and browser. In *Proceedings of the Fourth Papillon Workshop*, pages 40–48, Sapporo, Japan: Hokkaido University, 2003.

[SPS04] P. Smrž, M. Povolný, and A. Sinopalnikova. OASIS — A new tool for the transformation of XML knowledge resources into OWL. Poster at the Third International Semantic Web Conference, ISWC (http://iswc2004.semanticweb.org/posters/index.html), 2004.

[SS04] A. Sinopalnikova and P. Smrž. Corpus Analysis for Lexical Database Construction: A Case of Russian and Czech Wordnets. In *Proceedings of the Fourth International Conference on 33rd International Conference on Linguistics*, pages 23–29, Saint-Petersburg State University Press, Saint-Petersburg, Russia, 2004.

[SŽ98] P. Smrž and E. Žáčková. New tools for disambiguation of Czech texts. In *Proceedings of the First International Workshop on Text, Speech, Dialogue (TSD)*, pages 129–134, 1998. ISBN 80-210-1900-X.

Other Bibliography

[AMS+02] I. Azarova, O. Mitrofanova, A. Sinopalnikova, and I. Oparin. Building the lexical database for the Russian language. In *Proceedings of the Wordnet Structures and Standardization, and how these affect Wordnet Applications and Evaluation, LREC*, pages 60–64, ELRA, 2002.

[BDK+02] M. Butt, H. Dyvik, T. H. King, H. Masuichi, and C. Rohrer. The parallel grammar project. In *Proceedings of the Workshop on Grammar Engineering and Evaluation, COLING*, pages 1–7, Taipei Taiwan, 2002.

[BLHL01] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(5):34–43, May 2001.

[Bod99] R. Bod. *Beyond Grammar — An Experience-Based Theory of Language*. Cambridge University Press, Cambridge, 1999. ISBN 1-57586-151-8.

[Boo69] T. Booth. Probabilistic representation of formal languages. In *Tenth Annual IEEE Symposium on Switching and Automata Theory*, 1969.

[Bre01] J. Bresnan, editor. *Lexical-Functional Syntax*. Blackwell Publisher, Oxford, 2001.

[CHB+99] M. Collins, J. Hajič, E. Brill, L. Ramshaw, and C. Tillmann. A statistical parser of Czech. In *Proceedings of Thirty-Seventh ACL Conference*, pages 505–512. University of Maryland, College Park, USA, 1999.

[Chu04] K. Church. Speech and language processing: Can we use the past to predict the future? In P. Sojka, I. Kopeček, and K. Pala, editors, *Proceedings of the Seventh International Conference on Text, Speech and Dialogue (TSD)*, Lecture Notes in Artificial Intelligence LNCS/LNAI 3206, pages 3–13, 2004. ISBN 3-540-23049-1.

[CJ98] C. Chelba and F. Jelinek. Exploiting syntactic structure for language modeling. In C. Boitet and P. Whitelock, editors, *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, pages 225–231, San Francisco, California, 1998. Morgan Kaufmann Publishers.

[Fli00] D. P. Flickinger. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6(1):15–28, 2000.

[GP81] G. Gazdar and G. K. Pullum. Subcategorization, constituent order, and the notion “head”. In M. Moortgat, M.v.d. Hulst, and T. Hoekstra, editors, *The Scope of Lexical Rules*, pages 107–123. Foris, Dordrecht, Holland, 1981.

[Haj98] J. Hajič. Building a syntactically annotated corpus: The Prague dependency treebank. In *Issues of Valency and Meaning*, pages 106–132, Prague, 1998. Karolinum.

[Hor01] A. Horák. *The Normal Translation Algorithm in Transparent Intensional Logic for Czech*. PhD thesis, Faculty of Informatics, Masaryk University, Brno, 2001.

[HSS94] E. Hajičová, P. Sgall, and H. Skoumalová. An automatic procedure for topic-focus identification. *Computational Linguistics*, 21(1):81–94, 1994.

[Jel90] F. Jelinek. Self-organized language modeling for speech recognition. In A. Waibel and Kai-Fu Lee, editors, *Readings in Speech Recognition*. Morgan Kaufman Publishers, Inc., San Mateo, CA, 1990.

[Jos87] A. K. Joshi. An introduction to tree adjoining grammar. In A. Manaster-Ramer, editor, *Mathematics of Language*, pages 87–114. John Benjamins, 1987.

[JvN01] J.-C. Junqua and G. van Noord, editors. *Robustness in Language and Speech Technology*. Dordrecht: Kluwer Academic Publishers, 2001. ISBN 0-7923-6790-1.

[KM02] D. Klein and C. D. Manning. Natural language grammar induction using a constituent-context model. In T. G. Dietterich, S. Becker, and Z. Ghahramani, ed., *Advances in Neural Information Processing Systems 14, Cambridge, MA, MIT Press*, 2002.

[Kru01] G.-J. Kruiff. *A Categorial-Modal Logical Architecture of Informativity. Dependency Grammar Logic & Information Structure*. PhD thesis, Charles University, Prague, 2001.

[LGO04] LinGO Lab. CSLI linguistic grammars online, 2004. (`http://lingo.stanford.edu/`).

[MBF$^+$90] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Five papers on wordnet. Technical Report CSL Report 43, Cognitive Science Laboratory, Princeton University, 1990.

[MK93] J. T. Maxwell III and R. M. Kaplan. The interface between phrasal and functional constraints. *Computational Linguistics*, 19(4):571–590, 1993.

[MK00] S. Müller and W. Kasper. HPSG analysis of German. In W. Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, Arificial Intelligence, pages 238–253. Springer, Berlin, 2000.

[MSM94] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19(2):313–330, 1994.

[Nov03] V. Nováček. HPSG for Czech, 2003. Student's Project (IB030).

[Prz00] A. Przepiorkowski. Slavic languages in head-driven phrase structure grammar. `http://http://dach.ipipan.waw.pl/HPSG/slavic.html`, 2000.

[PS94] C. Pollard and I. A. Sag. *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago, 1994.

[Ric04] F. Richter. HPSG in Tuebingen, 2004. (`http://www.sfs.uni-tuebingen.de/hpsg/sysen.html`).

[SAJ88] Y. Schabes, A. Abeille, and A. K. Joshi. Parsing strategies with lexicalized grammars: Application to tree adjoining grammars. In *Proceedings of the Twelveth International Conference on Computational Linguistics (COLING'88)*, pages 578–583, Budapest, Hungary, 1988.

[Sed05] R. Sedláček. *Morphemic Analyser for Czech*. PhD thesis, Faculty of Informatics, Masaryk University, Brno, 2005.

[SHP86] P. Sgall, E. Hajičová, and J. Panevová. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecht, Netherlands, 1986.

[Sie00] M. Siegel. HPSG analysis of Japanese. In W. Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, Arificial Intelligence, pages 265–280. Springer, Berlin, 2000.

[SPH04] P. Sgall, J. Panevová, and E. Hajičová. Deep syntactic annotation: Tectogrammatical representation and beyond. In *HLT-NAACL Workshop on Frontiers in Corpus Annotation*, pages 32–38. Association for Computational Linguistics, 2004.

[TCS04] D. Tufis, D. Cristea, and S. Stamou. Balkanet: Aims, methods, results and perspectives. A general overview. *Romanian Journal of Information Science and Technology, Special Issue on the BalkaNet Project*, 7(9–43), 2004. (ISSN 1453-8245).

[Tic94] Tichý, P.: *The Analysis of Natural Language*, From the Logical Point of View III, 2:42–80, 1994

[Usz02] H. Uszkoreit. New chances for deep linguistic processing. In *Proceedings of the Nineteenth International Conference on Computational Linguistics (COLING)*, pages xiv–xxvii, Taipei, Taiwan, 2002. The Association for Computational Linguistics.

[Vos98] P. Vossen, editor. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht, 1998.

[XTAG01] XTAG Research Group. A lexicalized tree adjoining grammar for English. Technical Report IRCS-01-03, IRCS, University of Pennsylvania, 2001.

[Zip35] G. K. Zipf. *The Psycho-Biology of Language*. Boston, MA: Houghton Mifflin, 1935.

[ŽL04] Z. Žabokrtský and M. Lopatková. Valency frames of Czech verbs in VALLEX 1.0. In *HLT-NAACL Workshop on Frontiers in Corpus Annotation*, pages 70–77. Association for Computational Linguistics, 2004.

Chapter 2

A Parallel Metagrammar for Closely Related Languages — A Case Study of Czech and Russian

Smrž, P.: Parallel Metagrammar for Closely Related Languages — A Case Study of Czech and Russian. Accepted for publication in the Journal of Research on Language and Computation, Special Issue on Shared Representation in Multilingual Grammar Engineering, 27 pages, 2005.

Chapter 3

Grammatical Heads Optimized for Parsing and Their Comparison with Linguistic Intuition

Kadlec, V., Smrž, P.: Grammatical Heads Optimized for Parsing and Their Comparison with Linguistic Intuition. In Proceedings of the Seventh International Conference on Text, Speech and Dialogue, TSD. Berlin: Springer-Verlag, pp. 95–102, 2004.

Chapter 4

Probabilistic Head-Driven Chart Parsing of Czech Sentences

Smrž, P., Horák, A.: Probabilistic Head-Driven Chart Parsing of Czech Sentences. In Proceedings of TSD 2000. Berlin: Springer-Verlag, pp. 81–86. ISBN 3-540-41042-2, 2000.

Chapter 5

Efficient Sentence Parsing with Language Specific Features: A Case Study of Czech

Horák, A., Smrž, P.: Efficient Sentence Parsing with Language Specific Features:A Case Study of Czech. In Proceedings of the IWPT (the 7th International Workshop on Parsing Technologies). Beijing: Peking University, pp. 221–224, 2001.

Chapter 6

How Many Dots Are Really Needed for Head-Driven Chart Parsing?

Smrž, P., Kadlec, V.: How Many Dots Are Really Needed for Head-Driven Chart Parsing? Accepted for publication in Proceedings of Sofsem, 10 pages, 2006.

Chapter 7

Incremental Parser for Czech

Smrž, P., Kadlec, V.: Incremental Parser for Czech. In Proceedings of the 4th International Symposium on Information and Communication Technologies (WISICT05), Cape Town, South Africa, pp. 1–6, 2005.

Chapter 8

Determining Type of TIL Construction with Verb Valency Analyser

Smrž, P., Horák, A.: Determining Type of TIL Construction with Verb Valency Analyser. In Proceedings of SOFSEM'98. Berlin: Springer-Verlag, pp. 429–436. Lecture Notes in Computer Science 1521, ISBN 3-540-65260-4, 1998.

Chapter 9

A New Czech Morphological Analyser ajka

Sedláček, R., Smrž, P.: A New Czech Morphological Analyser ajka. In Proceedings of the 4th International Conference on Text, Speech and Dialogue (TSD). Berlin: Springer-Verlag, pp. 100–107. ISBN 3-540-42557-8, 2001.

Chapter 10

Building Czech Wordnet

Pala, K., Smrž, P.: Building Czech Wordnet. Romanian Journal of Information Science and Technology, Special Issue on the BalkaNet Project, Romanian Academy, 7, 1–2, pp. 79–88, ISSN 1453-8245, 2004.

Chapter 11

Lexical Databases as a Base for Broad Coverage Ontologies

Smrž, P.: Lexical Databases as a Base for Broad Coverage Ontologies. In Proceedings of the AAAI Workshop on Ontologies and the Semantic Web, AAAI. Edmonton, Alberta, Canada: AAAI Press, pp. 11–15, ISBN 1-57735-164-9, 2002.

Chapter 12

Quality Control and Checking for Wordnet Development: A Case Study of BalkaNet

Smrž, P.: Quality Control and Checking for Wordnet Development: A Case Study of BalkaNet. Romanian Journal of Information Science and Technology, Special Issue on the BalkaNet Project, Romanian Academy, 7, 1–2, pp. 173–182, ISSN 1453-8245, 2004.

Chapter 13

Lexical Databases in XML

Smrž, P., Povolný, M.: Lexical Databases in XML. In Proceedings of the EACL03 Workshop on Language Technology and the Semantic Web: The Third Workshop on NLP and XML (NLPXML-2003). Budapest, Hungary: ACL, pp. 49–55, 2003.

Chapter 14

Manatee, Bonito and Word Sketches for Czech

Rychlý, P., Smrž, P.: Manatee, Bonito and Word Sketches for Czech. In Proceedings of the Second International Conference on Corpus Linguistics. Saint-Petersburg: Saint-Petersburg State University Press, pp. 124–132, ISBN 5-288-03531-8, 2004.