Czech Science Foundation - Part C
Project Description

Applicant: doc. RNDr. Vladimír Petkevič, CSc.

Name of the Project: Syntactic annotation of Czech corpora

The current state of corpus tagging

Text corpora are an invaluable source of linguistic evidence, answering the need for richer and more representative data in virtually all linguistic disciplines. This need, aided by the development in computer technology, resulted in the creation of a whole new field – corpus linguistics. To better serve their purpose, corpora can be tagged, i.e. enhanced with some linguistic information: lemma, part of speech or a set of morphological categories.

Some corpora also include syntactic annotation, usually in the form of a syntactic tree, which is why such corpora are called treebanks. The oldest and best known treebank is the Penn Treebank (Marcus et al., 1993), built since 1989. It was tagged automatically, but corrected manually, with a significant amount of human effort. Since then, other treebanks were created for several tens of languages. As creating a treebank incurs high costs in terms of tedious manual work, the largest existing treebanks reach the relatively modest sizes of several million words, an insufficient number for many tasks.

The only large syntactically annotated corpus of Czech is the Prague Dependency Treebank (Hajič, 2006; PDT, 2006), built at the Faculty of Mathematics and Physics at Charles University. It contains about 1.5 million words, manually annotated with dependency trees, explicitly following the theoretical framework of Functional Generative Description (Sgall et al., 1986).[1]

The relatively small size of the Prague Dependency Treebank and its theoretical bias is a reason to further explore the path to a representative treebank of Czech. Among the morphologically tagged corpora, the Czech National Corpus (CNC, 2008) contains in its largest text resource about 500 million words. As the new millennium is approaching the end of its first decade, we believe CNC is ready for syntactic annotation. The Prague Dependency Treebank, developed with so much effort, has provided both the data needed for a practically usable parsing tool and the experience useful for making the further step. We plan to provide syntactic annotation for the Czech National Corpus with an unbiased annotation scheme, search interface using a representation accessible to lay users, and more options for experts.

[1] See Kučera (2006) and Hladká and Kučera (2008) for a project intended to make the Prague Dependency Treebank more accessible to students, where one of the main modifications is a different geometry of the syntactic tree, corresponding to the standard introduction to syntax as presented in Czech schools.

Goals of the project

Intuitive representation

Our project will result in procedures and tools for syntactic annotation of Czech texts, intended primarily for texts in the Czech National Corpus. Unlike the Prague Dependency Treebank, with its theoretically motivated design and primarily academic audience, a resource of this type has users beyond the research community, which imposes specific requirements on the annotation scheme and its representation in the corpus search interface. Syntactic information, including syntactic structure, should be easy to understand, following the common linguistic intuition and allowing for various amount of detail to be displayed even in a linear notation, with the tree graph display as an option and the rest of available information present in the data for other, more sophisticated uses. The default representation of syntactic structure should be more or less in line with the standard known to Czech students at the higher elementary and secondary levels, who are exposed to a relatively extensive introduction to syntax as a part of their curriculum. This system is largely based on Šmilauer (1966), inspired in turn by practical needs of primary and secondary schools, rather than by a particular syntactic theory. The concepts and the search interface should then be easy to understand for secondary school graduates without any special linguistic training. We expect that the annotated corpus will be useful to teachers of Czech language at all levels, to students and, indeed, to linguists.

Multiple interpretations

While presenting an easy, friendly interface to the lay user, the syntactic annotation scheme should not impose a single way of representing syntactic structure on everyone. The scheme should allow for multiple interpretations: in addition to the "easy", secondary-school-like format, there are at least two other obvious options: the dependency-based representation corresponding to the output of the stochastic parser, and a representation based on flat constituency trees. The last option will make the annotation more attractive to the international community of linguists, most of whom have no background in dependency grammar. We believe that the option to interpret the annotated data according to the user's (theoretical) preference will be an important and welcome feature of the search interface and a viable design principle of the annotation scheme.

Ambiguity and partial information

Corpus annotation is mostly unambiguous. Yet ambiguity is sometimes inevitable for fundamental reasons, whether on the level of morphology or syntax.[2] Additionally, unresolved ambiguity may be preferable to an arbitrary decision in case of poor evidence or some other insufficiency. For these reasons, our annotation scheme should allow for the possibility of preserving ambiguity by means of underspecification or (local) disjunction at all levels. In the most extreme case, a sentence can be annotated as a single syntactic chunk including a string of words. A partial analysis may identify a word's head, or its membership in a constituent, or its syntactic function, or any combination of the above, while still leaving other syntactic relationships in the sentence unresolved. We do not expect that unresolved ambiguity will be

[2]Examples include valency slots with ambiguous case requirements filled by nouns exhibiting case syncretism, or structures involving PP-attachment ambiguity without a difference in meaning (see for example, Oliva (2001)).

our preferred solution if unambiguous interpretation is attainable, but we wish to leave it as an option for all other cases.

Minimal human intervention

We wish to make full use of previous efforts and build on top of as many tools and resources as possible and practical. The most obvious candidates are tools developed using the Prague Dependency Treebank as their training data, especially a stochastic parser (Holan and Žabokrtský, 2006). The main benefit is that unlike most previous efforts, our syntactic annotation procedure can be automatic (except for the evaluation of results, leading to improvement and development of error correction tools). This will make the method suitable for annotating large corpora, such as the Czech National Corpus, where proofreading all annotated text for manual corrections is entirely unrealistic. Instead, an important part of the project will be evaluation of results and analysis of errors in the syntactic annotation. Results of the analysis will be used to build and optimize automatic error correcting tools.

Methodology

The source of data

The input data will be extracted from the Czech National Corpus (CNC, 2008) without linguistic tags. The text will be morphologically analyzed (using the analyzer described in Hajič (2004)) and then disambiguated. For morphological disambiguation, we will use a tool based on linguistic rules, developed at our institute (Petkevič, 2006), in combination with a stochastic tool morče (Votrubec (Raab), 2005).

Stochastic parsing and conversion of results

In the next step, the morphologically tagged texts will be parsed by a tool included in the tectomt package (Žabokrtský et al., 2008), namely by the stochastic parser developed by Holan and Žabokrtský (2006). Its success rate of 86% makes it currently the best performing parser of Czech.[3] The output consists of dependency trees, corresponding to the levels of surface and underlying syntax (*a-level*, *t-level*) of the Prague Dependency Treebank.[4] Syntactic structure, syntactic functions and other relevant information identified by the parser will be extracted from the PDT format and transformed into the new annotation scheme, designed for the purpose of presenting the annotated corpus to a wide linguistic public. Any other details of the annotation will be available on as an option to experts.

Annotation scheme and its representation

The new scheme abandons the strict dependency tree geometry of the source in favour of a less arbitrary and more intuitive treatment of potentially contentious phenomena: coordination, constructions including function words (verbal auxiliaries, prepositions, particles, subordinating conjunctions), and multi-word units, while allowing for the specification of dual heads (in

[3] This result is measured on PDT texts. The parser's error rate may drop by up to 2% for some type of texts.

[4] Both syntactic levels of PDT will be useful: only *t-level* includes explicit referential links.

some frameworks corresponding to a distinction between syntactic heads in surface and deep syntax), useful in many constructions mentioned above, and also in some constructions involving quantified nominals.[5] But all this does not mean that the original dependency format will be lost in the new scheme, nor that the scheme will impose a single mode of representing syntactic structure.

The envisaged format of the data will encode potentially multiple interpretations of a single string of text words, each of these interpretations consisting of three basic characteristics of syntactic structure: word order (including morphological markup), constituency (potentially discontinuous), and dependency (potentially multi-headed):

- The word-order dimension is a linear structure of (possibly underspecified or disjunctive) morphological tags referring to orthographical words in the text string, with the two structures usually corresponding in order and cardinality. The exceptions to the 1:1 pattern include mismatches between orthographical and syntactic words, represented, e.g., by the agglutinating clitic auxiliary *s*: *kdes*, or frozen multi-word units: *lážo plážo*.

- The constituency dimension consists of constituents or chunks, referring to items in the word-order structure or to other constituents. In a full parse of a sentence, it is equivalent to a flat phrase structure tree, but partial trees are not excluded.

- The dependency dimension consists of links referring to items in the word order structure. In a full parse of the sentence, it is equivalent to a dependency tree. Links can be labelled, possibly by syntactic functions, and can be of multiple types. This allows for deep and surface dependency relations to be expressed within one structure, or for the treatment of some constructions as multi-headed (coordination).

In the standard case, all links corresponding to the parse tree will be available: those relating heads with their dependents, as well as those grouping words into constituents. However, this setup may also support underspecification and disjunction at an arbitrary level, allowing for head-less or multi-head constituents, multiple or partial (sub)trees, or even a sequence of chunks interspersed with words as a very partial syntactic analysis. Multiple options may be treated as disjoint to represent ambiguity, or complementary to represent different dimensions, such as dependency links corresponding to surface and deep syntax.

Data in such a format will be searched and displayed through an interface capable of extracting and representing syntactic structure according to the user's preference. The default will be the "easy" mode, corresponding to the school standard, but other options will be available to access all information present in the annotated text in an efficient and intuitive way.

Improving the baseline success rate

To improve the success rate of the stochastic parser (our baseline), we plan the following measures:[6]

- The input to the stochastic parser is a fully disambiguated result of morphological analysis, with an error rate of its own (about 4%). The parser must cope with this initial

[5] In a treebank of Polish, Przepiórkowski (2007) distinguishes *syntactic* and *semantic* heads.

[6] Improvements will be measured on comparable data, i.e. excluding any cases involving underspecification or disjunction.

handicap as a given fact, and cannot backtrack to another, more plausible morphological tag, rejected in the previous step. We intend to experiment with multiple alternative options for a single word at the input to the parser and choose the optimal scenario.

- Steps following the stochastic parsing will be largely determined by the results of an extensive analysis of errors. The analysis will include evaluation of multiple parses (n-best) to find the optimal strategy for improving the parsing result, with the three complementary approaches:
  - Choosing the parse (sub)tree that looks most promising according to a metric (satisfying most / violating least linguistically motivated criteria)
  - Manipulating the best parse tree to correct an identified error
  - Introducing underspecification or local disjunction to express genuine ambiguity, or ambiguity as a solution to difficult problems

The set of tools used in this task represents a radical extension of our rule-based morphological disambiguation paradigm (Petkevič, 2006), including a syntax-oriented version of negative rules, designed to prune parse forests, and will be aided by the following lexically specific information:

- Valency lexicon (Skoumalová, 2001; Lopatková et al., 2008)
- Dictionary of multi-word units (Hnátková, 2006)
- Semantic classification of nouns (acquired from a thesaurus, e.g., WordNet (Pala and Smrž, 2004))

We expect that some improvements will require semantic information, and that the fuzzy nature of semantic regularities will find its proper implementation in stochastic methods targeted on individual lexemes.

Evaluation

For the purpose of evalution of our results we will use a balanced manually annotated corpus sample. The baseline will be the result of the stochastic parser achieved on the sample corpus (after conversion to our format). Then we will run the correcting procedures and compare the result with the manually annotated corpus again.

As an experiment, we will annotate a sample of the Prague Dependency Treebank texts manually according to our syntactic concept, and compare it with the real PDT data converted to our system of syntactic representation.

Schedule

In the first year of the project we will complete the following tasks:

- Designing annotation scheme and corpus search interface
- Testing available tools and their customization

- Creating data format converters
- Manually annotating a sample corpus (several tens of thousand words)

The second year:

- Implementing corpus search interface
- Automatic annotation of the sample corpus
- Error analysis, error classification
- Developing tools for automatic correction of errors
- Linguistic rules for the correction tools

The third year:

- Comparison of a converted PDT sample with our results
- Error analysis
- Improving linguistic rules
- Annotation of the entire experimental corpus (1 million words)
- Manual correction of results

Results and their presentation

The results of our project will be twofold – practical and theoretical; and journal articles and conference papers. The practical results will include

- Design of a syntactic annotation scheme supporting (i) a hybrid dependency/constituency format, and (ii) underspecified or disjunctive expressions at various levels of granularity
- A set of tools for syntactic annotation of Czech texts incorporating the stochastic parser, using the above scheme, with a success rate improvement on comparable data over the baseline results of the stochastic parser
- Pilot version of a corpus search interface capable of user-friendly interaction with the new syntactic annotation scheme
- Syntactically annotated experimental corpus
- In-depth analysis of parsing results, potentially useful for the development of more efficient methods and tools beyond the scope of the present project

Theoretical results:

- Evaluation of results of the stochastic parser, with a detailed qualitative analysis of errors

- Paradoxically, we believe that an effort to design a theoretically unbiased annotation scheme with multiple representation options should bring theoretical fruit.

As for the conference papers, we plan to present our work at the following conferences:

- Text, Speech and Dialogue (TSD)
- Language Resources (LREC)
- CLARET workshop
- (E)ACL conference
- Computational Linguistics – Applications Workshop (CLA)
- The International Workshop on Treebanks and Linguistic Theories

We also plan submissions to journals on computational or corpus linguistics, such as *International Journal of Corpus Linguistics*, *Prague Bulletin of Mathematical Linguistics*, and *Corpus Linguistics and Linguistic Theory*. A collection of papers on project-related topics can also be published in the series *Studie z korpusové lingvistiky* (Studies in Corpus Linguistics, publisher: Nakladatelství Lidové noviny, Prague).

Hardware and human resources

Our institute is well equipped with hardware: a server with 8 CPU's and another workgroup server equipped with 6 thin clients (Sun Ray). As for human resources, we expect that the work will be done by the following collaborators:

doc. RNDr. Vladimír Petkevič, CSc. – Head of the Institute of Theoretical and Computational Linguistics, an expert in the field of corpus linguistics, participated in many projects supported by GAČR or other institutions. Often participates in international projects; was the Czech team leader in projects CONCEDE (Consortium for Central European Dictionary Encoding; 1998–2000; PL-1142) and MULTEXT-EAST (Multilingual Text Tools and Corpora for Central and Eastern European Languages; 1995–1997; COP106). Also took part in projects Trans-European Language Resources Initiative (TELRI; 1995–1997) and Language Technologies for Slavic Languages (LATESLAV; PECO 2824; 1993–1995).

In the present project, he will be responsible for its management, and for the development of linguistic rules improving the output of the stochastic parser. Bonus: 80,000 CZK per annum.

RNDr. Milena Hnátková, CSc. – Researcher, member of the Institute of Theoretical and Computational Linguistics. Her main interests are morphology, phraseology, collocations, and linguistic rules for shallow parsing.

In the project, she will be responsible for identifying collocations, and she will also participate in error analysis and the development of linguistic rules. Bonus: 80,000 CZK per annum.

Mgr. Tomáš Jelínek – Researcher, part-time member of the Institute of Theoretical and Computational Linguistics and graduate student of mathematical linguistics. His main interests are syntax and development of linguistic rules for shallow parsing.

In the project, he will work on linguistic rules and error analysis. Project workload: 30%, we apply for 170,000 CZK per annum (incl. personal compensation) to extend his part-time workload from 70% to 100%.

Ing. Alexandr Rosen, Ph.D. – Researcher, deputy head of the Institute of Theoretical and Computational Linguistics. His main interests are formal grammars, linguistic theories, and representation of language structures.

In the project, he will be responsible for the design of the syntactic annotation scheme and the search interface, and he will also work on error analysis. Bonus: 80,000 CZK per annum.

RNDr. Hana Skoumalová, Ph.D. – Researcher, member of the Institute of Theoretical and Computational Linguistics. Her main interests are morphology, syntax and verb valency.

In the project, she will cooperate on the design of the syntactic representation and she will also work on linguistic rules and error analysis. Bonus: 80,000 CZK per annum.

Computational linguist / programmer – responsible for the implementation of (i) data conversion routines, (ii) parse trees re-ranking and error correction rules, and (iii) the corpus search interface. He will also cooperate on the design of the syntactic annotation scheme. We plan to hire a full-time collaborator with the salary of 400,000 CZK per annum (incl. personal compensation). The candidate must prove her experience with natural language processing and software development, have the degree Ph.D. (or equivalent), or be a graduate student of *mathematical linguistics* (or an equivalent discipline).

Our present candidate for this post is **RNDr. Jiří Hana, Ph.D.**, currently a Software Developer at the Center for Human Resource Research at The Ohio State University. He was involved in the morphological annotation of the Prague Dependency Treebank and of the Czech Academic Corpus, creating tools and guidelines for annotators. His main interests are morphology, tagging, and mathematical models of grammar. In case he will not be available at the moment of project start, we will hire another person who will meet our criteria.

Jiřina Kovaříková – Assistant at the Institute of Theoretical and Computational Linguistics. Will be responsible for administrative tasks, and will also work on error evaluation and auxiliary tasks. We apply for 55,000 CZK per year (incl. personal compensation) for her to extend her part-time workload of 50% at the faculty to 75 %.

In addition to these participants, we plan to hire students for the evaluation of data and other auxiliary tasks.

References

CNC (2001–2008). Czech National Corpus. `http://www.korpus.cz`. Institute of Czech National Corpus, Charles University, Faculty of Arts. Prague.

Gelbukh, A., editor (2007). *Computational Linguistics and Intelligent Text Processing (CICLing 2007)*, Lecture Notes in Computer Science, Berlin. Springer-Verlag.

Hajič, J. (2004). *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Karolinum, Charles University Press, Prague.

Hajič, J. (2006). Complex Corpus Annotation: The Prague Dependency Treebank. In Šimková, M., editor, *Insight into the Slovak and Czech Corpus Linguistics*, pages 54–73, Bratislava, Slovakia. Veda.

Hladká, B. and Kučera, O. (2008). An Annotated Corpus Outside its Original Context: A Corpus-Based Exercise Book. In *Proceedings of the ACL-08: HLT Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 36–43, Columbus, Ohio, USA. The Ohio State University.

Hnátková, M. (2006). Typy a povaha komponentů neslovesných frazémů z hlediska lexikálního obsazení. In Čermák, F. and Šulc, M., editors, *Kolokace*, volume 2 of *Studie z korpusové lingvistiky*, pages 142–167. Nakladatelství Lidové noviny, Prague.

Holan, T. and Žabokrtský, Z. (2006). Combining Czech Dependency Parsers. In *Proceedings of the 9th International Conference on Text, Speech and Dialogue*, volume 4188/2006, pages 95–102, Berlin/Heidelberg. Springer. ISBN 978-3-540-39090-9.

Kučera, O. (2006). Pražský závislostní korpus jako cvičebnice jazyka českého. Master's thesis, MFF UK Praha.

Lopatková, M., Žabokrtský, Z., and Kettnerová, V. (2008). *Valenční slovník českých sloves*. Karolinum, Charles Univeristy Press, Prague.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19.

Oliva, K. (2001). On retaining ambiguity in disambiguated corpora. *TAL (Traitement Automatique des Langues)*, 42(2).

Pala, K. and Smrž, P. (2004). Building Czech Wordnet. *Romanian Journal of Information Science and Technology*, 7(1-2):79–88. ISSN 1453-8245.

PDT (2006). Prague Dependency Treebank. `http://ufal.mff.cuni.cz/pdt2.0/`. Institute of Formal and Applied Linguistics, Charles University, Faculty of Mathematics and Physics. Prague.

Petkevič, V. (2006). Reliable Morphological Disambiguation of Czech: Rule-Based Approach is Necessary. In Šimková, M., editor, *Insight into the Slovak and Czech Corpus Linguistics*, pages 26–44. Veda, Bratislava. ISBN 80-224-0880-8.

Przepiórkowski, A. (2008). *Powierzchniowe przetwarzanie języka polskiego*. Akademicka Oficyna Wydawnicza EXIT, Warszawa.

Przepiórkowski, A. (2007). On heads and coordination in valence acquisition. In Gelbukh (2007), pages 50–61.

Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Academia/D. Reidel, Prague/Dordrecht.

Skoumalová, H. (2001). *Czech syntactic lexicon*. PhD thesis, Charles University, Faculty of Arts, Prague.

Votrubec (Raab), J. (2005). Volba vhodné sady rysů pro morfologické značkování češtiny. Master's thesis, MFF UK, Prague.

Šmilauer, V. (1966). *Novočeská skladba*. Státní pedagogické nakladatelství, Praha, 3 edition.

Žabokrtský, Z., Ptáček, J., and Pajas, P. (2008). TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer. In *Proceedings of the 3rd Workshop on Statistical Machine Translation, ACL*.

PartC - Attachment

Proposal of foreign experts

1. **doc. Adam Przepiórkowski, PhD**

   field: computational linguistics: deep and shallow parsing of Polish
   corpus linguistics
   information extraction
   machine learning methods in NLP

   e-mail: adamp@ipipan.waw.pl

   address: Polish Academy of Sciences
   Institute of Computer Science
   ul. Ordona 21
   01-237 Warszawa
   Poland

   url: http://nlp.ipipan.waw.pl/

2. **Dr. Roland Meyer**

   field: Slavic linguistics and corpus linguistics
   Czech corpus-base grammar
   formal grammar of Slavic languages

   e-mail: roland_meyer@sprachlit.uni-regensburg.de

   address: Altes Finanzamt
   Institut für Slavistik
   Universitätsstraße 27
   93040 Regensburg
   Germany

   url: http://www.uni-regensburg.de/Fakultaeten/phil_Fak_IV/Slavistik/

3. **Prof. Dr. Tilman Berger**

   field: corpus linguistics
   Slavic linguistics

   e-mail: tberger@uni-tuebingen.de

   address: Slavisches Seminar der Universität Tübingen
   Wilhelmstraße 50
   72074 Tübingen
   Germany

   url: http://www.slavistik.uni-tuebingen.de/

Czech Science Foundation - Part D
Applicant and Co-applicants

Applicant: doc. RNDr. Vladimír Petkevič, CSc.

Curriculum Vitae

1974-79 studied at Charles University, Faculty of Mathematics and Physics

1985 was awarded the title RNDr. (approx. MSc.)

1979-92 worked in the Research institute of mathematical machines

1992 defended disertation in the field of *computer science and informatics*

1993-now works at Charles University, Faculty of Arts

1994-now director of the Institute of Theoretical and Computational Linguistics at the Faculty of Arts

1996 habilitation in the field *mathematical lingustics* with the work *Underlying Structure of Sentence Based on Dependency*

1996-now head of the synchronic linguistic section of the Institute of Czech National Corpus, Faculty of Arts

Research and grants

2003-05 GAČR 405/03/0377 *Možnosti a meze gramatiky češtiny ve světle Českého národního korpusu* (Exploring the Core and Limits of Czech Grammar as seen through the Czech National Corpus); head investigator: F. Štícha – rated as excellent

2003-05 GAČR 405/03/0086 *Slovní poklad češtiny v informační společnosti* (The Word Thesaurus of Czech in the Information Society); head investigator: J. Králík – rated as excellent

2003-05 GAČR 405/03/0913 *Velké jazykové korpusy a jejich automatická analýza* (Very Large Language Corpora and Their Automatic Analysis); head investigator: J. Hajič – rated as excellent

2004-07 GA AV ČR 1ET100610409 *Diagnostické a evaluační nástroje pro lingvistický software* (Diagnostic and Evaluation Tools for Linguistic Software); head investigator: K. Oliva

2005-11 Research Project MSM0021620823 *Český národní korpus a korpusy dalších jazyků* (Czech National Corpus and Corpora of Other Languages); head investigator: F. Čermák

Selected bibliography

Petkevič, V. (1995a). A new formal specification of underlying structures. *Theoretical Linguistics*, 21(1):7–61. ISSN 0301-4428. 3 citations on WoS, 1 citation on Google Scholar.

Petkevič, V. (1995b). *Underlying Structure of Sentence Based on Dependency*. Charles University, Faculty of Arts, Prague. Habilitation work. 2 citations on Google Scholar.

Petkevič, V. (2006a). Reliable morphological disambiguation of czech: Rule-based approach is necessary. In Šimková, M., editor, *Insight into the Slovak and Czech Corpus Linguistics*, pages 26–44. Veda, Bratislava. ISBN 80-224-0880-8. No citations found.

Petkevič, V. (2006b). Složité předložkové skupiny (kolokace předložek a jmen). In Čermák, F. and Šulc, M., editors, *Studie z korpusové lingvistiky II. Korpusová lingvistika: Kolokace*, pages 262–310. Nakladatelství Lidové noviny / Ústav Českého národního korpusu. ISBN 80-7106-863-2. No citations found.

Petkevič, V. (2008). Structure of the nominal group in the czech national corpus and its part-of-speech and morphological disambiguation. In et al., G. Z., editor, *Formal Description of Slavic Languages: The Fifth Conference, Leipzig 2003*, pages 53–67, Frankfurt am Main. Peter Lang. ISSN 1436-6150, ISBN 978-3-631-55160-8. No citations found.

Petkevič, V. and Tláskal, J., editors (2005). *Josef Vachek: Lingvistický slovník Pražské školy*. Nakladatelství Karolinum, Prague. Translation. ISBN 80-246-0933-9. 4 citations on Google Scholar.

Čermák, F., Klímová, J., and Petkevič, V., editors (2000). *Studie z korpusové lingvistiky*. Nakladatelství Karolinum, Prague. ISBN 80-7184-893-X. ISSN 0567-8269. 4 citations on Google Scholar.

Čermák, F. and Petkevič, V. (2005). Linguistically motivated tagging as a base for a corpus-based grammar. In Danielsson, P. and Wagenmakers, M., editors, *Proceedings of Corpus Linguistics 2005*, The Corpus Linguistics Conference Series, Birmingham. `http://www.corpus.bham.ac.uk/PCLC/#grammar`. ISSN 1747-9398. No citations found.

Results since 2004

B (monograph): 3

C (chapter in a book): 2

D (article in proceedings): 9

T (prototype, verified technology, software product, ...): 1

S (prototype, methodology, functional sample, authorized software, ...): 1

A(V) (accessed remotely): 1

Number of citations on Web of Science

30

H-index

2