Czech Science Foundation - Part C
Project Description

Applicant: doc. RNDr. Markéta Lopatková, Ph.D.

Name of the Project: *Delving Deeper: Lexicographic Description of Syntactic and Semantic Properties of Czech Verbs*

Information on syntactic and semantic properties of verbs, which are traditionally considered to be the center of a sentence, plays a key role in many rule-based NLP tasks such as information retrieval, text summarization, question answering, machine translation, etc. Moreover, theoretical aspects of valency and its adequate lexicographic representation are also challenging.

The proposed project focuses on a theoretically adequate description of advanced language phenomena related to the valency behavior of verbs. Changes in valency structure of verbs belong to such phenomena – whereas various types of changes in valency structure of verbs, based on different language means (grammatical, syntactic or semantic), attract the attention of linguists for decades (see Section 1), their adequate lexicographic representation is still missing. For studying such verbal properties, a concept of a (type) situation – consisting of a set of participants that are characterized by particular semantic properties and connected by particular relations – plays a key role. Thus, mapping participants onto valency complementations and their semantic characteristics are crucial for a lexicographic processing of language data.

The project takes advantage of an existing Valency lexicon of Czech verbs VALLEX. VALLEX provides the core valency information, i.e. information on the combinatorial potential of verbs in their particular senses. The goals of the project are two fold. It aims at the deepening of the theoretical insight into various phenomena at the syntax-semantics interface (including the contrastive perspective). Moreover, it attempts to substantially qualitatively and quantitatively enhance the publicly available electronic language resource.

1 Background of the Project

The relation of syntactic and semantic properties of verbs belongs to the topics which have been widely discussed in recent times among a broad community of researchers. This project primarily focuses on two research areas: (i) on studying the changes in valency structure of verbs, and (ii) on a possibility of mutual enrichment of language resources storing different types of syntactic and semantic information.

In Czech linguistics, the study of syntactic constructions characterized by changes in valency structure of verbs from the syntactic point of view started in the late sixties, mainly under the influence of Russian linguistics, esp. (Apresjan, 1967; Mel´čuk, Cholodovič, 1970; Apresjan, 1974; Chrakovskij, 1977), recently esp. (Padučeva, 2002; Chrakovskij, 2005; Boguslavskij, 2008). The terms hierarchization, diathesis or conversion were introduced in Czech and Slovak grammars, see esp. (Daneš, 1968; 1985) and also (Ondrejovič, 1989; Štícha, 1984; Grepl, Karlík, 1998). Roughly speaking, such terms refer to changes in mutual assignment of semantic participants and (surface) syntactic positions, while the real situation expressed by sentences remains the same.

In American linguistics, there are four basic approaches for representing changes in valency structure of verbs, (i) structurally based approaches represented mainly by transformational-generative grammars, esp. (Chomsky, 1957; 1965; Baker, 1988), (ii) lexically based approaches specifying all alternating valency frames in a lexicon, esp. (Bresnan, 1982), (iii) correspondence rules approach formulating extralexical rules establishing the mapping between participants and their surface syntactic positions (Jackendoff, 1990, Bresnan, 2001), and (iv) constructionally based approaches based on the assumption that a difference in syntactic forms marks a difference in meaning, esp. (Borer, 2005; Goldberg, 1995).

There is a number of language resources providing information on syntactic and semantic properties of verbs. Let us mention at least the most significant language resources for English that store deep syntactic information and that represent a highly relevant source of information for Czech language resources.

FrameNet[1] is an on-line lexical database documenting semantic and syntactic combinatory possibilities (valences) of each word in each of its senses (Baker et al., 1998). FrameNet is based on frame semantics (Fillmore et al., 2003) and supported by corpus evidence: each lexical unit (a pair consisting of a word and its meaning) evokes a particular semantic frame underlying its meaning. Each semantic frame is conceived as a "conceptual structure describing a particular type of situation, object, or event", (Ruppenhofer et al., 2006); it contains the so-called frame elements, i.e., semantic participants of such situations. The FrameNet database contains more than 10 thousand lexical units in more than 825 semantic frames, exemplified by more than 135 thousand annotated sentences.

P. Hanks and J. Pustejovsky proposed the method of so-called **Corpus Pattern Analysis**[2]. It is a technique that offers a systematic analysis of the patterns of meaning and use of each verb (rather than specification of the set of its separate meanings), based on a large sample of its corpus utterances, see esp. (Hanks, Pustejovsky, 2005). The valences of verbs are analyzed and semantic types and semantic roles are assigned to each valence. A semantic type is an intrinsic property of a valence of lexical unit, like [Person], [PhysObj], [Concept]. By contrast, a semantic role is context-specific and it is assigned by the context of a verb occurrence (e.g., Doctor or Patient in medical treatment context). The method of CPA is used as a basic approach in building a **Pattern Dictionary of English**[3], PDEV (in prep., see Hanks, 2008).

'Levin's classes'. A significant contribution to the examination of relations between syntactic and semantic properties of English verbs was made by B. Levin (Levin, 1993) who introduced the term 'alternation' referring to changes in valency structure of English verbs. Based on an extensive list of various types of alternations, she suggested a rich system of semantic classes. Although her work deals with English verbs, a lot of parallels with syntactic behavior of Czech verbs can be found there. On the basis of Levin's classes, rich semantic classification of verbs is built in the VerbNet project.

As a theoretical background of the proposed project, the **Functional Generative Description** (henceforth FGD) is adopted, see esp. (Sgall et al., 1986). In FGD, valency is related primarily to the tectogrammatical layer, i.e., the layer of linguistically structured meaning. The valency characteristics are encoded in a form of a valency frame, which is modeled as a sequence of frame slots corresponding to valency complementations of a verb labeled by (rather coarse-grained) tectogrammatical roles such as 'Actor', 'Patient', 'Effect', 'Direction', etc., see esp. (Panevová, 1994). In addition, possible morphemic forms are specified for each valency complementation.

The valency theory of FGD was applied to a large amount of data within the **Prague Dependency Treebank**[4], PDT 2.0. Moreover, this theory is used for tectogrammatical annotation of the large parallel **Prague Czech-English Dependency Treebank**, PCEDT (in preparation, manual annotation) and **CzEng** (automatic annotation)[5]. The FGD valency theory also serves as a framework for the valency lexicon of Czech verbs VALLEX 2.5 (Lopatková et al., 2008).

VALLEX 2.5[6] provides information on the valency structure of verbs in their particular senses: on the number of valency complementations, on their type labeled by functors, and on their morphemic forms, (Žabokrtský, Lopatková, 2007; Lopatková et al., 2008). VALLEX 2.5 describes 2730 verb lexemes containing about 6460 lexical units typically corresponding to one sense. At present, more

[1] http://framenet.icsi.berkeley.edu/

[2] http://nlp.fi.muni.cz/projects/cpa/

[3] http://deb.fi.muni.cz/pdev/

[4] http://ufal.mff.cuni.cz/pdt2.0/

[5] http://ufal.mff.cuni.cz/czeng/

[6] http://ufal.mff.cuni.cz/vallex/2.5/

than 44% of lexical units are divided into heterogeneous 'supergroups' − as 'communication', 'mental action', 'motion', 'exchange', 'transport', etc. −, which represent a rather tentative classification, based primarily on similar morphosyntactic patterns (number of valency complementations, their morphemic forms and specific syntactic properties) and similar semantics.

Let us mention other lexical resources providing information on valency behavior of Czech verbs. These works represent additional valuable sources of language data for us.

VerbaLex (esp. Hlaváčková, Horák 2006; Hlaváčková 2008) provides information on syntactic and semantic features of Czech verbs, including semantic verb classes (based on the VerbNet project) and semantic roles / deep cases (based on the 1st-Order-Entity and 2nd-Order-Entity basis from the EuroWordNet Top Ontology and on the Princeton WordNet Base Concepts). The VerbaLex lexicon is linked to the Czech WordNet semantic network.

Czech Syntactic Lexicon (Skoumalová, 2011) represents another electronic resource for Czech. It contains valency frames of around 15 000 Czech verbs, the lexicon entries provide also information on reflexivity, some types of grammatical diatheses, and other syntactic characteristics of verbs. This automatically created lexicon is primarily designed for the purposes of NLP.

Finally, there are also two available printed Czech valency lexicons: *Slovesa pro praxi*, see (Svozilová et al., 1997) and *Slovník slovesných, substantivních a adjektivních vazeb a spojení*, see (Svozilová et al., 2005).

The above mentioned language resources represent extremely valuable sources of Czech, English as well as Czech-English parallel data which are essential for any theoretical research. These resources containing syntactic and/or semantic information will be complemented by extensive morphologically annotated corpora of various languages (esp. Czech National Corpus[7], Russian National Corpus[8], IPI PAN Corpus[9], PWN Corpus of Polish[10] and other parallel corpora[11]) and **word sketches**, an automatic corpus-derived summary of a word’s grammatical and collocational behavior[12].

2 Description of the Project

The main goal of the proponed project is to propose an adequate framework for the theoretical description of language phenomena at the syntax-semantics interface; such a framework will be applied in lexicographic processing of language data. A close interplay between theoretical research and its application to an extensive data annotation represents a fruitful strategy that fortifies both sides involved.

The preliminary investigation leads us to the identification of the following areas which should be addressed in the project:

1 The lexicographic representation of various changes in valency structure of verbs

- 1.1 Theoretical research; design of a formal model for lexicographic description
- 1.2 Grammatical and syntactic diatheses: theoretical and practical aspects
- 1.3 Semantic diatheses: theoretical and practical aspects
- 1.4 Other types of changes in valency structure of verbs
- 1.5 Comparative aspects of diatheses
- 1.6 Application in an electronic language resource

2 Mapping lexical resources: an effective way of enriching lexical information

- 2.1 Enhancing Czech valency lexicon with semantic classes and semantic roles
- 2.2 Strengthening lexical resources with corpus evidence

2.1 The lexicographic representation of various changes in valency structure of verbs

Syntactic behavior of verbs is determined to a great extent by their lexical semantic properties. Prototypically, a single meaning of a verb corresponds to a single valency structure. However, in many

[7] http://ucnk.ff.cuni.cz
[8] http://www.ruscorpora.ru
[9] http://korpus.pl
[10] http://korpus.pwn.pl
[11] Available through the website of the Institute of Czech Ntional Corpus, http://ucnk.ff.cuni.cz
[12] http://www.korpus.cz/corpora/

cases semantically related uses of verbs can be syntactically structured in different ways. E.g., the pairs of sentences in (1a)-(1b), (1a)-(2a) and (1b)-(2b) differ in their syntactic structure despite their obvious semantic similarity:

(1) a. *Petr naložil vůz senem. // Peter loaded the truck with hay.*
b. *Petr naložil seno na vůz. // Peter loaded hay on the truck.*

(2) a. *Vůz byl naložen senem. // The truck was loaded with hay.*
b. *Seno bylo naloženo na vůz. // Hay was loaded on the truck.*

Here we use the term **diathesis** for specific relations between uses of the same verb lexeme: these uses exhibit semantic affinity (i.e., they describe the same situational content that consists of a set of **situational participants** characterized by particular semantic properties and connected by certain relations); however, they are syntactically structured in different ways. Instead of the term diathesis, the term **alternation** is sometimes used as a most general term grouping similar changes in syntactic-semantic structure of verbs (see Levin, 1993; Levin, Rappaport Hovav, 2005).

Although thoroughly discussed in theoretical works for many languages (see Section 1), a question remains how to describe changes in syntactic structure of verbs associated with diatheses in a valency lexicon. Listing separate entries for each of them makes the lexicon bigger than expected; more importantly, such a massive polysemy of verbs seems to be contra-intuitive.

2.1.1 Theoretical research; design of a formal model for lexicographic description

The results already achieved (see Section 1) together with the previous studies carried out by the project team (esp. Kettnerová, Lopatková 2009a; Kettnerová, Lopatková 2010a) form a solid basis for further research. These studies have revealed that the available electronic resources (together with comparative approach, see Section 2.1.5 below) can bring new facts that either are not covered in the classical studies or do not conform to existing descriptions (e.g., so called diatheses of the second instance *Když se dostane přidělena nová pracovna.* are not supported by corpus evidence).

In principle, we can distinguish three types of Czech diatheses according to the linguistic means that express them: (i) grammatical diatheses (subtask 1.2), (ii) syntactic diathesis (subtask 1.2) and (iii) semantic diatheses (subtask 1.3). Under certain conditions, these types can be combined together.

2.1.2 Grammatical and syntactic diatheses: theoretical and practical aspects

We can distinguish several types of changes expressed by the grammatical means – passive, deagentive, resultative, recipient passive and dispositional diathesis, see esp. (Skoumalová, 2001; Kettnerová, Lopatková, 2009a; Panevová, in print; Panevová et al., manuscript). Based on our previous research, we consider these types as regular enough to be captured by formal syntactic rules. These syntactic rules should be stored in the grammar component of the lexicon. In the data component, there should be a single lexical unit representing both uses of a verb.

The same representation seems to be adequate also for those changes in valency structure that are expressed by syntactic means – esp. reciprocity plays a central role here (Panevová, 1999; 2007).

The proposed project focuses on further theoretical research; the main goal is to formulate detailed syntactic rules describing the concerned phenomena.

2.1.3 Semantic diatheses: theoretical and practical aspects

Semantic diathesis is a specific relation between two (or more) uses of the same lexeme that are characterized by the same situational content (usually consisting of the same set of situational participants with the same semantic properties and relations among them). However, these constructions show different syntactic structures. As a prototypical semantic diathesis, so called Locatum-Location diathesis is commonly cited (example (1a)-(1b) above, more detailed description is provided in Daneš, 1685; Kettnerová, Lopatková, 2010a). In contrast to grammatical and syntactic diatheses, changes in valency frames associated with semantic diatheses exhibit many irregularities in their valency characteristics (concerning a type of valency complementations, their morphemic form(s) as well as obligatoriness).

Moreover, a slight semantic shift between such syntactic constructions shows evidence for separate lexical units. In the data component, these units should be represented by separate entries and interlinked by a general rule specifying the relevant type of semantic diathesis; such rules should be stored in the grammar component, see esp. (Kettnerová, Lopatková, 2010a).

The proposed project focuses on the delimitation of relevant language phenomena (based on existing works both for Czech and other languages), on their detailed formal description and on their representation in the valency lexicon.

2.1.4 Other types of changes in valency structure of verbs

The preliminary investigation has revealed also other types of constructions that are characterized by the changes in valency structure of Czech verbs (Kettnerová, Lopatková, 2010b), namely multiple structural expression of a situational participant (as in example (5)) and structural splitting of a situational participant (6):

(5) a. *Turisté vylezli na kopec. // Tourists climbed up the hill.*
b. *Turisté vylezli kopec. // Tourists climbed the hill.*
(6) a. *Petr řekl, že je Marie chytrá. // Peter said that Mary was clever.*
b. *Petr řekl o Marii, že je chytrá. // 'Peter-said-about-Mary-that-(she) is-clever.'*

We plan to study similar constructions in the Czech language having in mind their appropriate lexicographic description.

2.1.5 Comparative aspects of diatheses

Some language phenomena, relatively rare in one language, can belong to core features of another language and thus are intensively studied. Since diatheses rank among such phenomena, a comparative point of view brings a better insight into the problem and contributes to the full description of these phenomena.

Comparative research of diatheses has a long tradition – Leningrad typology school described some diatheses in detail for different languages, see esp. (Kategorija zaloga, 1970; Chrakovskij, 1974). However, the linguists of this school only rarely examined other Slavic languages than Russian. In the comparative research of Slavic languages, linguists have paid the closest attention to particular grammatical diatheses (Havránek, 1928-1937; partly Doros, 1975; Wieczorek, 1982; Běličová-Křížková, 1976; Běličová, Uhlířová, 1996; Grepl, 1973).

A comparison of Russian and both Sorbian languages is provided in (Jermakova, 1992); (Topolińska, 1984) provides a comparison of Polish, Serbo-Croatian and Macedonian; the authors attempt to give a complete overview of diatheses in their articles.

Here we focus especially on a comparison of Czech, Russian and Polish on the background of other Slavic languages. There are a lot of interesting questions in comparative research such as the existence of particular diatheses in observed languages, frequency of unmarked vs. marked member of diathesis, possibility of expression (incl. possible morphemic forms) of individual participants with respect to the language typology etc. Comparative point of view gives us the possibility of a more exact description of Czech diatheses.

2.1.6 Application in an electronic language resource

The results of the research carried out in the previous subtasks are going to be applied in the VALLEX lexicon. This will either involve a development of automatic data-driven procedures identifying possible diatheses (esp. for grammatical diatheses; this method is advocated esp. in Skoumalová, 2001) or a manual annotation of individual lexical entries (esp. for semantic diatheses).

The main output of this subtask is the valency lexicon VALLEX enhanced (i) with the grammar component specifying especially grammatical aspects of individual diatheses and (ii) with the information on applicability of individual diatheses for individual lexical units in the data component.

2.2 Mapping lexical resources: an effective way of enriching lexical information

The second research area of the proposed project focuses on the enhancement of the lexicon VALLEX with a syntactico-semantic classification of verbs and on an exploitation of this classification. Further, adding corpus evidence for individual lexical units represents an important source of knowledge which can be used especially for NLP tools.

2.2.1 Enhancing Czech valency lexicon with semantic classes and semantic roles

A number of well-grounded language resources provide information on semantic roles. However, they employ different theoretical frameworks whose theoretical assumptions are reflected in annotation schemes. This has an important consequence: each lexical resource captures different types of information. Linking information from several lexical resources then represents an effective way of enriching a particular lexical resource.

On the other hand, differences in theoretical assumptions beyond lexical resources bring several difficulties with mapping information: the different level of granularity in word sense disambiguation represents a typical example. Moreover, other requirements for harmonizing linguistic information are imposed on interlinking information from different-lingual lexical resources: especially an accurate translation represents a fundamental prerequisite for successful mapping.

Based on our preliminary studies, we have decided to use the FrameNet lexical database (see above). In a pilot project (Kettnerová et al., 2008; Kettnerová, Lopatková, 2009b), the verbs from chosen semantic 'superclasses' (as they are implemented in the current version of VALLEX) were classified into more coherent semantic classes based on semantic frames from FrameNet. Then we assigned frame elements as semantic roles to each valency complementation of given verbs. The proposed classification exploits the principles of the 'Inheritance' relation defined in FrameNet data – prototypically, semantic frames from appropriate upper levels of abstraction are used as semantic classes. Similar technique can be used for semantic roles classification – as suitable semantic role labels, we propose to use frame elements that belong to semantic frames selected as semantic classes.

The proposed method allows us to overcome the problem with the different level of granularity made in the resources. The inter-annotator agreement attained in the pilot project was satisfactory (e.g., assigning semantic frames for verbs of exchange: IAA 78.5%, κ measure 0.73; assigning frame elements for verbs of exchange: IAA 91.2%, κ 0.91, see Kettnerová at al., 2009b). The results of the pilot phase of the project have proved the feasibility of the proposed method of enriching lexical information provided in VALLEX with the information from FrameNet.

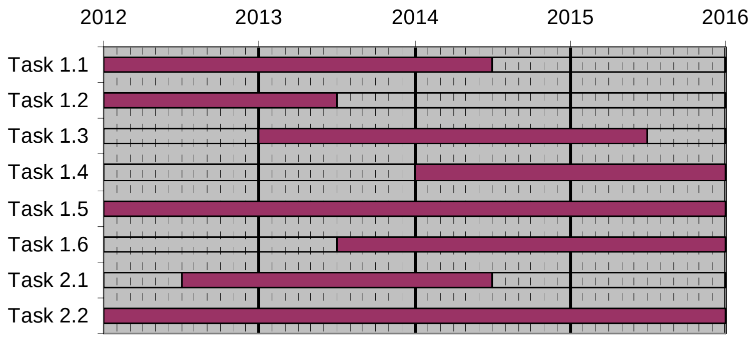
Though the proposed method has already provided us with interesting theoretical results for selected groups of verbs, many interesting questions still remain open (especially concerning verbs that do not belong to a distinctive semantic class). Further, the exploitation of the proposed method for the lexicographic work, as well as its massive application in the VALLEX lexicon, belong to the topics of the presented project.

2.2.2 Strengthening the Lexical Resource with Corpus Evidence

Corpus evidence is an important source of information for many NLP tools. For VALLEX, version 1.0, the VALEVAL golden data – consisting of a sample of 10 256 corpus sentences with (manually) assigned lexical units to the occurrences of 109 Czech verbs – are available. The PDT-VALLEX (Hajič et al., 2003; Urešová, 2006; 2010) and EngVallex lexicons (Cinková, 2006; Bojar, Šindlerová, 2010) are both linked to sentences from PDT and PEDT, respectively. Similarly, the PDEV lexicon (Hanks, 2008) exploits the evidence from the British National Corpus.

The goal of this minor subtask of the project is to (semi-)automatically merge information from the available Czech and English resources through the existing lexicons (esp. PDT-VALLEX, EngVallex and PDEV lexicons). A universal methodology and schema suitable for linking external lexicographic resources have been designed (Bejček et al., 2010); the proposed data format overcomes different logical structures of lexicons. Moreover, this format allows for an effective way how to visualize and search the data.

2.3 Project Time Table



3 Methods used

The proposed project takes advantage of the combination of four basic methodological approaches.

3.1 The first one is the traditional method of collecting linguistic evidence, its classification and exploitation. This approach is unavoidable for formulating any appropriate theoretical description of language phenomena; moreover, the role of thorough linguistic evidence even increases when one deals with deep syntactic and semantic analysis. This method is grounded on the available corpora and lexicons (see Section 1) and methods of their advanced searching (including, e.g., word sketches). It also exploits a great amount of investigation and knowledge contained in various publications (see Sections 1, 2).

3.2 The second one is a method of formulating a suitable formal framework for the description of separate syntactic and semantic phenomena as well as of their interplay. Here, the possibility of testing proposed solutions plays a crucial role. The project follows the tradition of FGD and its emphasis on the adequacy and the economy of descriptions; further, testable criteria and applicability of proposed solutions to language data are an essential requirement.

3.3 The third method consists in an application of the theoretical conclusions to a large amount of data. This approach makes it possible to verify theoretical conclusions, to identify theoretical assumptions that deserve more thorough explanation as well as to reveal possible inadequate postulates.

3.4 When creating language resources, a thorough testing and evaluation of data consistency, as well as an inter-annotator agreement on individual language phenomena belong to standard requirements in a NLP community.

4 The goals of the project and their practical exploitation

The topic of the project is highly relevant; at present, many linguists, focusing on the description of higher layers of languages, i.e., deep (underlying) syntax and semantics, concentrate their attention on the theoretical description of the language phenomena at the syntax-semantics interface (i.e., not only on the central valency phenomena, which have been studied since the middle of the last century, but also on valency phenomena belonging to the boundary between the center and the periphery of the language as well as on purely peripheral phenomena). Moreover, they focus on capturing valency behavior of particular lexemes in dictionaries.

The description and representation of syntactic and semantic properties of verbs are not only a concern of theoretical linguistics: valency lexicons cannot be ignored by advanced applications in Natural Language Processing (NLP) which are based on the explicit description of language (often denoted as 'rule-based' approaches) either. At the same time they are necessary for building language resources which are used by NLP tools based on machine learning ('data-driven' approaches).

5 The expected outputs of the project

The proposed project will have the following main outputs:

5.1 Theoretical and methodological results of the project will be published in journals dedicated to Czech and other Slavic languages (esp. categories J_{imp} and J_{neimp}), in thematic anthologies and as chapters of monographs (category C). In addition, these results will be presented at the international and Czech conferences of both theoretical and computational linguistics (especially at those with proceedings monitored in WoS, category D).

5.2 The significant output of the project consists in developing a framework for mutual (semi)automatic enhancement of various language resources; such framework includes data formats and algorithms for interlinking individual resources; moreover, (semi)automatic procedures for enhancing lexicons with corpus data are planned.

5.3 The main applied output of the project is both a qualitatively and quantitatively enhanced valency lexicon of Czech verbs available for a wide professional audience as well as for students and other language users. An emphasis will be laid on both human and machine-readability; thus both linguists and developers of applications within the Natural Language Processing domain can use it.

The lexicon will be published both as a monograph (category B) and as an electronic language resource (software, category R).

6 Research team

The applicant and her co-workers employ their long term experience with formal description of the Czech language. Semantic and syntactic properties of Czech verbs belong to their core research interests and this is also reflected in an ample list of publications on valency and related topics (see also form D1). They also have extensive experience with various applied tasks related to the proposed project, such as lexicographic description of valency (VALLEX) or complex annotation of Czech data (Prague Dependency Treebank). These activities clearly prove their interest and preparedness for the tasks involved in the proposed project.

The applicant Markéta Lopatková coordinates the project. She is one of the main authors of the concept of the valency lexicon VALLEX and of the methodology used to build it. She will be responsible for the close interplay between theoretical research and its application in the lexicon. The co-applicant Karolína Skwarska is a Slavic studies scholar with extensive experience with comparative syntactic studies (esp. Czech, Russian, Polish and Slovene). Václava Kettnerová is an advanced PhD student of computational linguistics with a master degree in Czech studies. They have already shown their competence in building the lexicon. The research team is completed by two PhD students of computational linguistics, Eduard Bejček (master degree in computer science) and Anna Lauschmannová (master degree in logic and mathematics); they will responsible for the design and implementation of suitable data formats as well as for the design and implementation of automatic and semiautomatic data processing.

The team also profits from the long-term cooperation with prof. Patrice Pognan from INALCO (Institut National des Langues et Civilisations Orientales) and LaLIC (Language, Logiques, Informatique, Cognition) at Paris-Sorbonne University (l'Université Paris-Sorbonne). Their recent collaboration focuses – among others – on the possibility of exploiting the VALLEX lexicon in a bilingual lexicon (ECO-NET project).

The Institute of Formal and Applied Linguistics has a number of undergraduate and postgraduate students, whose background is either in computer science or linguistics. These students may help with well-defined linguistic tasks as well as with automatic processing the data.

The Institute of Slavonic Studies co-operates with students and Ph.D. students of Slavonic studies at the Faculty of Arts, Charles University. The Institute would like to broaden this cooperation to the field of comparative syntax. Students should participate in processing the material from corpora and in other auxiliary works.

Both institutions – the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics and the Institute of Slavonic Studies of ASCR – are relatively well equipped with both hardware and software. The project will be able to profit from the availability of a computer network, printers, copying machines etc.

7 Bibliography

Apresjan, Ju. D. (1967) *Eksperimental'noe issledovanie semantiki russkogo glagola.* Moskva: Nauka.

Apresjan, Ju. D. (1974) *Leksičeskaja semantika. Sinonimičeskie sredstva jazyka*. Moskva: Nauka.

Baker, M. C. (1988) *Incorporation: A theory of grammatical function changing.* Chicago, IL: University of Chicago Press.

Baker, C. F., Fillmore, Ch. J., Lowe, J. B. (1998) The Berkeley FrameNet Project. In *Proceedings of the COLING-ACL*, Montreal, Canada.

Běličová-Křížková, H. (1976) Kategorie osoby a systém diateze v slovanských jazycích. Ke vztahu morfologické a syntaktické roviny v jazyce. *Slavia*, Vol. 45, p. 337-355.

Běličová, H., Uhlířová, L. (1996) *Slovanská věta*, Praha: Euroslavica.

Bejček, E., Kettnerová, V., Lopatková, M. (2010) Advanced Searching in the Valency Lexicons Using PML-TQ Search Engine. In *Proceedings of the Text, Speech, Dialog Conference 2010*, LNCS 6231, Berlin / Heidelberg: Springer, p. 51-58.

Boguslavskij, I. M. (2008) Tol'ko li u glagolov jest' diatezy? *Voprosy jazykoznanija*, No. 6, p. 6-28.

Bojar, O., Šindlerová, J. (2010) Building a Bilingual ValLex Using Treebank Token Alignment: First Observations. In *Proceedings of the Language Resources and Evaluation Conference 2010*, Paris: ELRA, p. 304-309

Borer, H. (2005) *The Normal Course of Events*. Oxford: Oxford University Press.

Bresnan, J., editor (1982) *The Mental Representation of Grammatical Relations*. Cambridge, Massachusetts / London, England: MIT Press.

Bresnan, J. (2001) *Lexical-Functional Syntax*. Oxford, Blackwell.

Cholodovič, A. A., editor (1974) *Tipologija passivnych konstrukcij. Diatezy i zalogi.* Leningrad: Nauka.

Chomsky, N. A. (1957) *Syntactic Structures*. The Hague: Mouton.

Chomsky, N. A. (1965). *Aspects of the Theory of Syntax*. Cambridge: MIT Press.

Chrakovskij, V. S., editor (1974) *Problemy lingvističeskoj tipologii i struktury jazyka*. Leningrad: Nauka.

Chrakovskij, V. S. (2005) The Concept of Diatheses and Voices in Early XXI Century. In *Vostok – Zapad. Vtoraja meždunarodnaja konferencija po modeli "Smysl – Tekst"*, Moskva: Jazyki slavjanskoj kul'tury, p. 529-537.

Cinková, S. (2006) From PropBank to EngValLex: Adapting the PropBank-Lexicon to the Valency Theory of the Functional Generative Description. In *Proceedings of the Language Resources and Evaluation Conference 2006*, Paris: ELRA, p. 2170-2175, 2006

Daneš, F. (1968) Some Thoughts on the Semantic Structure of the Sentence. *Lingua*, Vol. 21, p. 55-69.

Daneš, F. (1985) *Věta a text: studie ze syntaxe současné češtiny*. Praha: Academia.

Doros, A. (1975) *Werbalne konstrukcje bezosobowe w języku rosyjskim i polskim na tle innych języków słowiańskich*, Wrocław: Zakład Narodowy im. Ossolińskich.

Fillmore, C. J., Johnson, C., Petruck, M. R. L. (2003) Background to FrameNet. *International Journal of Lexicography*, Vol. 16, No. 3, p. 235-250.

Goldberg, A. E. (1995) *Constructions. A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

Grepl, M. (1973) Deagentnost a pasívum v slovanských jazycích. In *Československé přednášky pro VII. mezinárodní sjezd slavistů ve Varšavě,* Praha: Academia, p. 141-149.

Grepl, M., Karlík, P. (1998) *Skladba češtiny*. Olomouc: Votobia.

Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V., Pajas, P. (2003) PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In *Proceedings of the Treebanks and Linguistic Theories Workshop 2003,* Vaxjo: University Press, p. 57-68.

Hanks, P. (2008) Mapping Meaning onto Use: a Pattern Dictionary of English Verbs. In Abstracts of the *American Association of Corpus Linguistics Conference* 2008, Utah: Brigham Young University.

Hanks, P., Pustejovsky, J. (2005) A Pattern Dictionary for Natural Language Processing. *Revue Française de Langue Appliquée*, Vol. 10, No. 2, p. 63-82.

Hlaváčková, D. (2008) *Databáze slovesných valenčních rámců VerbaLex.* PhD Thesis, Masaryk University in Brno.

Hlaváčková, D., Horák, A. (2006) VerbaLex – New Comprehensive Lexicon of Verb Valencies for Czech. In *Computer Treatment of Slavic and East European Languages*, Bratislava, Slovakia: Slovenský národný korpus, p. 107-115.

Havránek, B. (1928, 1937) *Genera verbi ve slovanských jazycích*. Díl I., II. Praha: Královská česká společnost nauk. Díl

Jackendoff, R. S. (1990) *Semantic Structures*. Cambridge: MIT Press.

Kategorija zaloga (1970) Materialy konferencii. Leningrad: Nauka.

Kettnerová, V., Lopatková, M. (2009a): Changes in Valency Structure of Verbs: Grammar vs. Lexicon. In Levická, J., Garabík, R. (eds.) *Proceedings of Slovko 2009, NLP, Corpus Linguistics, Corpus Based Grammar Research*, Bratislava, Slovakia: Slovenská akadémia vied, p. 198-210.

Kettnerová, V., Lopatková, M. (2009b) *Mapping Semantic Information from FrameNet onto VALLEX.* Contributed talk, FrameNet Masterclass and Workshop, co-located with TLT 8, Milan.

Kettnerová, V., Lopatková, M. (2010a) The Representation of Diatheses in the Valency Lexicon of Czech Verbs. In *Proceedings of the Conference on Advances in Natural Language Processing (IceTAL 2010)*, LNCS 6233, Berlin / Heidelberg: Springer, p. 185-196.

Kettnerová, V., Lopatková, M. (2010b) Representation of Changes in Valency Structure of Verbs in the Valency Lexicon of Czech Verbs. In *Proceedings of Verb 2010, Interdisciplinary Workshop on Verbs, The Identification and Representation of Verb Features*, Pisa, Italy: Scuola Normale Superiore – Laboratore di Linguistica, Universita di Pisa – Dipartimento di Linguistica, p. 154-159.

Kettnerová, V., Lopatková, M., Hrstková, K. (2008) Semantic Classes in Czech Valency Lexicon: Verbs of Communication and Verbs of Exchange. In *Proceedings of the Tect, Speech, Dialog Conference 2008*, LNCS 5246, Berlin / Heidelberg: Springer, p. 109-116.

Jermakova, M. I. (1992) Tak nazyvaemye prjamye *genera verbi* v serbolužickich i russkom literaturnych jazykach. In Grek-Pabisowa, I., Smirnow, L. N. (eds.) *Synchroniczne badania porównawcze systemów gramatycznych języków słowiańskich*, Warszawa: Slawist. Ośrodek Wydawniczy, p. 45-56.

Levin, B. C. (1993) *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago and London: The University of Chicago Press.

Levin, B. C., Rappaport Hovav, M. (2005) Argument Realization. Cambridge: Cambridge University Press.

Lopatková, M., Žabokrtský, Z., and Kettnerová, V. (2008). *Valenční slovník českých sloves*. Nakladatelství Karolinum, Praha.

Mel'čuk, I. A. (1998) *Kurs obščej morfologii*, t. II. Moskva – Vena.

Mel'čuk, I. A., Cholodovič, A. A. (1970) K teorii grammatičeskogo zaloga. *Narody Azii i Afriki*, No. 4, p. 111-124.

Ondrejovič, S. (1989). *Medzi slovesom a vetou*. Jazykovedné štúdie. Bratislava: Veda.

Padučeva, Je. V. (2002) Diateza i diatetičeskij sdvig. In *Russian Linguistics*, Vol. 26, No. 2, p. 179-215.

Panevová, J. (1994) Valency Frames and the Meaning of the Sentence. In Luelsdorff, P. A. (ed.) The Prague School of Structural and Functional Linguistics, Amsterdam: John Benjamins Publishing Company, p. 223-243.

Panevová, J. (in print) O kategorii rezul'tativa (prežde vsego) v češskom jazyke (to appear In *Zbornik Matice srpske za slavistiku*)

Panevová, J. et. al. (manuscript) *Syntax současné češtiny (na základě anotovaného korpusu)*. Praha: Nakladatelství Karolinum.

Panevová, J. (1999) Česká reciproční zájmena a slovesná valence. *Slovo a slovesnost*, Vol. 90, p. 269-275.

Panevová, J. (2007) Znovu o reciprocitě. *Slovo a slovesnost*, Vol. 68, p. 91-100.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C., Schefczyk, J. (2006) *FrameNet II: Extended Theory and Practice*. http://framenet.icsi.berkeley.edu/book/book.html/

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

Skoumalová, H. (2001) *Czech Syntactic Lexicon.* PhD thesis, Charles University in Prague.

Skwarska, K. (in print) Semantičeskie diatezy v češskom jazyke s učetom russkogo i pol'skogo jazykov (to appear In *Zbornik Matice srpska za slavistiku*)

Svozilová, N., Prouzová, H., Jirsová, A. (1997) *Slovesa pro praxi.* Praha: Academia.

Svozilová, N., Prouzová, H., Jirsová, A. (2005) *Slovník slovesných, substantivních a adjektivních vazeb a spojení*. Praha: Academia.

Štícha, F. (1984) *Utváření a hierarchizace struktury větného znaku*. Praha: Univerzita Karlova.

Topolińska, Z. (1984) Właściwości diatetyczne czasowników w języku polskim, macedońskim i serbsko-chorwackim (założenia opisu typologicznego). *Studia konfrontatywne polskopołudniowosłowiańskie*, Wrocław: Zakład Nar. im. Ossolińskich, p. 103-134.

Urešová, Z. (2006): The Verbal Valency in the Prague Dependency Treebank from the Annotator's Point of View. In Šimková, M. (ed.) *Insight into Slovak and Czech Corpus Linguistic*s, Bratislava: Veda, p. 93-112.

Urešová, Z. (in press) Building the PDT-VALLEX valency lexicon. In *Proceedings of the Fifth Corpus Linguistics Conference*, Liverpool: UK.

Wieczorek, D. (1982) *Nesohlasovannye asimmetričnye russkie predloženia v sopostavlenii s pol'skimi*, Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego.

Žabokrtský, Z., Lopatková, M. (2007) Valency Information in VALLEX 2.0: Logical Structure of the Lexicon. *The Prague Bulletin of Mathematical Linguistics*, No. 87, p. 41-60.