

- a) postup prací při řešení grantového projektu/ rozbor řešení grantového projektu na pracovišti řešitele a spoluřešitelů včetně zdůvodnění případného odklonu řešení projektu proti schválenému návrhu projektu, dosažené výsledky projektu (u mezinárodních projektů včetně mezinárodní spolupráce):

In the fourth year of the project, all research areas were surveyed in accordance with the original proposal of the project. The accomplishments of the project are reflected in two journal articles (one in a journal with an impact factor [15-1], one in Jneimp category [15-10]). The achievements of the project were also presented at two international conferences (focused on formal linguistics and natural language processing) and published in their conference proceedings [15-2], [15-3] and [15-4]. A PhD thesis closely related to the central topics of the project (Automatické propojování lexikografických zdrojů a korpusových dat. [Automatic linking of lexicographic sources and corpus data]) was finished, submitted and successfully defended by one of the team members, Eduard Bejček [15-5]. In addition, three oral presentations were delivered [15-7], [15-8] and [15-9]. A main applied output of the project, the enriched on-line Valency Lexicon, VALLEX, version 3.0, was released [15-6] (see also <http://ufal.mff.cuni.cz/vallex/3.0/>).

In addition to these already presented and/or published outputs, the team members have prepared two more articles [00-2] and [00-3] (accepted for publication). Further, four texts [00-5], [00-6], [00-7] and [00-8] were prepared based on the presentations at the conference Výzkum slovesné valence ve slovanských zemích včera a dnes [Research of Verbal Valency in Slavic Countries in the Past and at Present] held by Slovanský ústav AV ČR, v.v.i. / Instytut Slawistyki in December 2014 (accepted for publication as chapters in a collective monograph).

The research conducted within the span of the whole project was summed up in the new printed edition of the lexicon Valenční slovník českých sloves [Valency lexicon of Czech verbs] VALLEX [00-4] (submitted to the Karolinum Press). The research teams of the collaborating institutions – the Institute of Formal and Applied Linguistics (MFF UK) and the Slavonic Institute (SI CAS) – closely collaborated: an internal seminar was organized in June, attended by all team members as well as by other researchers; research plans, partial results as well as open problems and possible solutions were intensively discussed via e-mail.

1 Lexicographic representation of various changes in valency structure of verbs

1.1 Theoretical research; design of a formal model of lexicographic description

MFF UK, SI CAS: Theoretical research into changes in valency structure of Czech verbs was completed. Based on final annotations, minor aspects of the formal model of lexicographic representation of Czech verbs were settled – a focus was laid esp. on syntactic reflexivity [15-1], on deletions [15-2], [15-3], on an interplay between the grammar and data components of the lexicon in case of complex predicates [15-4], and on differences between formal expression of verb complements in the trilingual lexicon [15-9].

These aspects were projected to the data component and incorporated in the VALLEX lexicon (as described in the following sections).

1.2 Grammatical and syntactic diatheses: theoretical and practical aspects

MFF UK: The work on Czech diatheses was completed (due to the extent of the necessary manual annotations, the work on the annotation of diatheses has taken longer than planned in the project proposal). A semi-automatic method identifying lexical units of Czech verbs that form marked constructions of individual diatheses has been proposed [00-6]. This method has been applied to the corpus data; the automatic processing was followed by a manual annotation. Moreover, the rules describing individual Czech diatheses were tested, refined, and evaluated with respect to the corpus data – these rules form the grammatical component of the lexicon, which is available both as machine readable file with syntactic rules and as part of the book edition of VALLEX [00-4], where it has been extended with explanatory examples. A special attention has been paid to the rules describing the surface form of the marked members of diatheses of verbs with a genitive complement [15-6], [00-6] (submitted also as [00-9]). Further, an interplay of the formal rules for diatheses and rules determining surface structure formation of complex predicates with light verbs was studied [15-4].

Moreover, all the theoretical results were incorporated in the lexicon data [15-6], [00-4].

SI CAS: The theoretical comparative research on diatheses (Czech-Russian-Polish) continued. A comparison of grammaticalized alternations, namely reciprocity, in Russian and Polish with respect to their Czech counterparts was presented [15-8], [00-2]. Moreover, reflexive constructions in Polish and Czech and resultative constructions in Russian and Czech were compared, [00-8] and [00-3], respectively. As for practical aspects, grammaticalized alternations of Russian and Polish lexical units (counterparts of relevant Czech lexical units) were annotated. Further, reflexive constructions with the accusative case specific of Polish have been annotated.

1.3 Semantic diatheses: theoretical and practical aspects

MFF UK: In the last year of the project, work on all three types of lexicalized alternations (= semantic diatheses) identified in the Czech language was carried out. Lexical rules describing changes in the correspondence between situational participants and valency complementations of verbs were refined and captured in the grammatical component of the lexicon [00-4] and [15-6]. Hereby all results of the theoretical research in the previous years were

applied to the lexicon data.

SI CAS: The annotation of Russian and Polish lexical units – counterparts of Czech lexical units of verbs entering lexicalized alternations in VALLEX 2.7 – was completed. Special attention was devoted to lexical-semantic ‘subject – object’ conversions in Russian and Polish both from the theoretical and from the practical point of view.

1.4 Other types of changes in valency structure of verbs

MFF UK: In the last year of the project, a corpus study of verbs involving verb complement was conducted – the study focused on changes in valency structure brought about by syntactic reflexivity [15-1]. Further, a lexicographic description of these changes was proposed [15-6], [00-4].

SI CAS: The annotation of relevant Polish and Russian lexical units has been finished.

1.5 Comparative aspects of diatheses

MFF UK and (mainly) SI CAS: All of the above mentioned types of alternations (diatheses, reciprocity, lexicalized alternations) were studied as described above. Further, the annotation of new groups of Russian and Polish verbs was carried out by means of the electronic tool prepared at MFF UK (attributes: ‘synon’, ‘example’, ‘control’, ‘aspect’, ‘diat’, ‘reciprocity’, ‘refl’, ‘split’, ‘conv’, ‘multiple’). Evidence from the Russian National Corpus (www.ruscorpora.ru), from the Polish National Corpus (www.nkjp.pl) and from Internet texts was used and corrected by a native speaker. The data were compared with the new version of VALLEX 3.0.

1.6 Application in an electronic lexical resource

The final version of the on-line Czech valency lexicon, VALLEX, version 3.0, was released [15-6]. Both data and grammar component were enhanced with the information on diatheses, reciprocity and lexical alternations. A large set of automatic checks was run in order to prevent any inconsistencies; further, an extensive manual checking of data consistency was conducted.

VALLEX is available through brand-new web interface <http://ufal.mff.cuni.cz/vallex/3.0/>. A printed edition has been prepared – it has been submitted to the Karolinum Press (the manuscript is attached; now under the reviewing process by three independent reviewers) [00-4].

SI CAS: The model of a trilingual electronic annotation tool for Russian and Polish verbs was proposed, refined and implemented (new attributes, linking of Czech-Russian-Polish counterparts etc.) in a close cooperation with E. Bejček (MFF UK).

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

2.1 Enhancing Czech valency lexicon with semantic classes and semantic roles

The effort aiming to linking the VALLEX lexicon with FrameNet continued as planned. The approach proposed was presented in [15-5]. However, the results of the developed automatic method, which would enable us to automatically add the missing semantic information (semantic classes and semantic roles), were not satisfactory enough; thus this information is not included in the final version of VALLEX 3.0.

2.2 Strengthening the lexical resource with corpus evidence

The VALLEX lexicon is enriched with corpus evidence from the Prague Dependency Treebank (PDT) and (the Czech part of) the Prague Czech-English Dependency Treebank (PCEDT) via interlinking VALLEX and the corpus based PDT-Vallex lexicon [15-5]. As a result, more than 2.5 thousand lexical units in VALLEX are enriched with a total of 75 thousand examples extracted from the PDT and PCEDT corpora, available in the on-line version of the lexicon [15-6]. An approach preferring high precision (reaching 91%) over coverage ensures high reliability of the obtained results. The automatic links to PDT-Vallex on the level of individual lexical units were also exploited to add examples of the marked members of the passive diathesis for some verbs. Additional corpus examples were selected during the manual stage of the annotation [00-6].

(Since the CzEngVallex – an external resource, not part of this project – was published as late as in October 2015, the information obtained from this lexicon could not be included in VALLEX as was planned in the project proposal.)

The amount of research conducted by the MFF UK as well as SI CAS research teams fully corresponds to the percentage proportion of the partners on the project.

Bibliography 2015:

[15-1] Kettnerová Václava, Lopatková Markéta, Panevová Jarmila: Shoda doplňku v reflexivních konstrukcích v češtině. Slovo a slovesnost, Vol. 76, No. 3, pp. 198-214, 2015.

[15-2] Hajič Jan, Hajičová Eva, Mikulová Marie, Mírovský Jiří, Panevová Jarmila, Zeman Daniel: Deletions and node reconstructions in a dependency-based multilevel annotation scheme. In CICLing 2015: 16th International Conference on Computational Linguistics and Intelligent Text Processing, LNCS 9041 Springer, Berlin / Heidelberg, pp. 17-31, 2015.

- [15-3] Hajičová Eva, Mikulová Marie, Panevová Jarmila: Reconstruction of Deletions in a Dependency-based Description of Czech: Selected Issues. In Depling 2015: Proceedings of the Third International Conference on Dependency Linguistics, Uppsala University, Uppsala, Sweden, pp. 131-140, 2015.
- [15-4] Kettnerová Václava, Lopatková Markéta: At the Lexicon-Grammar Interface: The Case of Complex Predicates in the Functional Generative Description. In Depling 2015: Proceedings of the Third International Conference on Dependency Linguistics, Uppsala University, Uppsala, Sweden, pp. 191-200, 2015.
- [15-5] Bejček Eduard: Automatické propojování lexikografických zdrojů a korpusových dat. Ph.D. thesis, Charles University in Prague, Faculty of Mathematics and Physics, Prague, Czech Republic, 227 pp., 2015.
- [15-6] Lopatková Markéta, Kettnerová Václava, Bejček Eduard, Vernerová Anna, Žabokrtský Zdeněk: VALLEX 3.0 - Valenční slovník českých sloves. Data/software, Charles University in Prague, Faculty of Mathematics and Physics, <http://ufal.mff.cuni.cz/vallex/3.0/>, 2015.
- [15-7] Bejček Eduard: Propojování valenčních slovníků. Oral presentation, Seminář formální a aplikované lingvistiky, 19. 10. 2015, ÚFAL MFF UK, Prague, Czech Republic, 2015.
- [15-8] Skwarska Karolína: Reciprocita v češtině, ruštině a polštině. Zpracování ve valenčním slovníku. Oral presentation. Moskva, 2015.
- [15-9] Skwarska Karolína: Где кончается синтаксис и начинается семантика при описании свойств славянского глагола. Oral presentation, Congress ICCEES Makuhari, 2015.
- [15-10] Skwarska Karolína.: Opis leksykograficzny syntaktycznych i semantycznych cech czasowników w języku polskim i rosyjskim na podstawie korpusów narodowych. Prace Filologiczne, Vol. LXVII, pp. 349-366, 2015.

Prepared for publication:

- [00-2] Skwarska Karolína: Reciprocita v češtině, ruštině a polštině. Zpracování ve valenčním slovníku. Slavia. Časopis pro slovanskou filologii. (accepted)
- [00-3] Podobová Marie: Bezobjektové posesivní rezultativní konstrukce od sloves tranzitivních a bezobjektové i objektové posesivní rezultativní konstrukce od sloves intransitivních – jejich tvorba a funkce v českém jazyce. Nová čeština doma a ve světě. (accepted)
- [00-4] Lopatková Markéta, Kettnerová Václava, Bejček Eduard, Vernerová Anna, Žabokrtský Zdeněk: Valenční slovník českých sloves VALLEX. (submitted to Karolinum)
- [00-5] Kettnerová Václava: Syntaktická struktura komplexních predikátů a její popis ve valenčním slovníku. In Skwarska, K., Kaczmarska, E. (eds.) Výzkum slovesné valence ve slovanských zemích. (accepted)
- [00-6] Lopatková Markéta, Vernerová Anna, Kettnerová Václava: Diateze ve Valenčním slovníku českých sloves VALLEX. In Skwarska, K., Kaczmarska, E. (eds.) Výzkum slovesné valence ve slovanských zemích. (accepted)
- [00-7] Panevová Jarmila: Valence v gramatice, valence ve slovníku. In Skwarska, K., Kaczmarska, E. (eds.) Výzkum slovesné valence ve slovanských zemích. (accepted)
- [00-8] Skwarska Karolína, Bolbot Katarzyna, Filas Adriana: Zpracování polských zvrtných konstrukcí ve slovníku VALLEX. In Skwarska, K., Kaczmarska, E. (eds.) Výzkum slovesné valence ve slovanských zemích. (accepted)
- [00-9] Vernerová Anna, Lopatková Markéta, Kettnerová Václava: Diateze u českých sloves s genitivním doplněním. Výzkum slovesné valence ve slovanských zemích včera a dnes 2016. (submitted)

b) využití technického a přístrojového vybavení pořízeného z grantových prostředků:

MFF UK: The costs on computer equipment renewal (PC components) were partially covered by the grant budget in accordance with the original project proposal.

SI CAS: No computer equipment was covered by the grant budget in 2015.

c) stručná zpráva o průběhu a výsledcích zahraničních cest:

-- DepLing 2015: International Conference on Dependency Linguistics, Uppsala, Sweden, August 24-26 2015 (<http://depling.org/depling2015/>); Markéta Lopatková (total 27.2 thousand CZK: air ticket, accommodation, conference fee, travel allowance); Václava Kettnerová (total 4.5 thousand CZK: air ticket); paper presentation [15-4];
-- ITAT 2015: Informačné Technológie – Aplikácia a Teória; workshop SloNLP 2015: Slovenskočeský NLP workshop (SloNLP 2015), Spišská Nová Ves, Slovakia September 17-21 2015; Markéta Lopatková (total 6.4 thousand CZK: train ticket, conference fee); consultations, member of the program committee;
-- ICCEES IX World Congress 2015: Makuhari, Japan, August 3-8 2015 (<http://www.l.u-tokyo.ac.jp/makuhari2015/index.html>); Karolina Skwarska (total 40.7 thousand CZK: air ticket, accommodation, travel allowance); oral presentation [15-9];
-- Meeting of the Commission on the Grammatical Structure of the Slavic Languages of the International Committee of Slavists, Moscow, Russia September 18-19 2015; Karolina Skwarska (total 17.1 thousand CZK: air ticket, accommodation, travel allowance); presentation [15-8].

d) celkové shrnutí řešení projektu (u mezinárodních projektů včetně mezinárodní spolupráce):

The accomplishments of the four year project Delving Deeper: Lexicographic Description of Syntactic and Semantic Properties of Czech Verbs are twofold:

1. The theoretical insight into various language phenomena at the syntactic-semantic interface has been deepened – both grammaticalized alternations (diatheses and reciprocity) and lexicalized alternations (lexical-semantic conversions, structural splitting of a situational participant and multiple structural realization of a situational participant) were put under scrutiny. A contrastive perspective was also applied: we have focused namely on differences in syntactic behavior of Czech, Polish and Russian verbs undergoing changes in their valency structure. Moreover, a typologically new type of syntactic change – related to syntactic reflexivity in Czech – was identified and studied in detail, esp. changes in morphemic expressions of verbal complementations conditioned by the long and clitic variant of the reflexive pronoun.

Based on these theoretical achievements, the formal model of the lexicon has been refined and enriched – as a result, the final model provides an adequate and economic lexicographic representation of the studied phenomena. Further, elaborate lexicographic rules for describing changes in valency behavior of verbs undergoing alternations have been designed.

2. The electronic Valency lexicon of Czech verbs VALLEX has been substantially qualitatively and quantitatively enhanced in two dimensions. First, the information on the applicability of individual alternations (both grammaticalized and lexicalized) was added to its data component. Moreover, the lexicon has been enriched with the sample annotations of Polish and Russian lexical units undergoing alternations. Second, VALLEX has been interlinked with the annotation lexicon of the Prague Dependency Treebank; as a result of the interlinking, the lexicon has been enriched with examples from the treebank. The lexicon is accessible on the following webpage <http://ufal.mff.cuni.cz/vallex/3.0/>.

The results were made available to the research community in journals dedicated to Czech and other Slavic languages (esp. categories Jimp, Jneimp and Jrec; 6 articles already published and 2 accepted for publication), in thematic anthologies and as chapters of monographs (category C; 1 published, 4 accepted for publication) and 1 theoretical monograph (already published). In addition, these results were presented at international and Czech conferences on both theoretical and computational linguistics (especially at those with proceedings monitored in WoS, category D; 5 published texts).

The main applied output of the project is both 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 was 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 is prepared for publication as a monograph (category B) and it has been already released as an electronic language resource (software, category R).

Another positive aspect of the project was the involvement of students, which resulted in 1 MSc thesis (Gregoire Labbé) and 2 PhD theses (Václava Kettnerová and Eduard Bejček). Five more PhD students also participated in the project (Anna Vernerová, Marie Podobová, Natalia Klyueva, Katarzyna Vaculová and Adriana Filas) and thus have obtained important research experience.