Towards a Corpus-based Valency Lexicon of Czech Nouns

Jana Klímová, Veronika Kolářová, Anna Vernerová

Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské nám. 25, 118 00 Prague 1, Czech Republic
E-mail: {klimova, kolarova, vernerova}@ufal.mff.cuni.cz

Abstract

Corpus-based Valency Lexicon of Czech Nouns is a starting project picking up the threads of our previous work on nominal valency. It builds upon solid theoretical foundations of the theory of valency developed within the Functional Generative Description. In this paper, we describe the ways of treating valency of nouns in a modern corpus-based lexicon, available as machine readable data in a format suitable for NLP applications, and report on the limitations that the most commonly used corpus interfaces provide to the research of nominal valency. The linguistic material is extracted from the Prague Dependency Treebank, the synchronic written part of the Czech National Corpus, and Araneum Bohemicum. We will utilize lexicographic software and partially also data developed for the valency lexicon PDT-Vallex but the treatment of entries will be more exhaustive, for example, in the coverage of senses and in the semantic classification added to selected lexical units (meanings). The main criteria for including nouns in the lexicon will be semantic class membership and the complexity of valency patterns. Valency of nouns will be captured in the form of valency frames, enumeration of all possible combinations of adnominal participants, and corpus examples.

Keywords: corpus, lexicon, nouns, valency

1. Introduction

Nominalizations as reclassifications of their corresponding verbal clauses (Heyvaert, 2003) are in the center of attention of many researchers; both their syntactic and semantic aspects are studied across various languages and frameworks (Chomsky, 1970; Osenova, 2009; Alexiadou & Rathert, 2010; Melloni, 2011). One such aspect is argument structure (Grimshaw, 1991) or nominal "valency" (Spevak, 2014): the number, type and form of arguments that are bound to a noun. Although nominal valency still remains in the shadow of the valency of verbs, it is the matter of both theoretical and lexicographic studies which are in a close relationship. This relationship may be best exemplified by the *Explanatory Combinatorial Dictionary of Modern Russian* (Mel'čuk & Zholkovsky, 1984), a dictionary created within the theoretical framework of the Meaning-Text Theory. While many valency lexicons are primarily intended for non-native speakers (e.g., Herbst et al., 2004) is intended for learners of English and the oldest lexicon covering nouns (Sommerfeldt & Schreiber, 1977) for learners of German), nouns are also covered in lexicons created mainly with NLP applications in mind, such as FrameNet[1] (ongoing; see also Ruppenhofer et al., 2006) and NomBank 1.0[2] (see also Meyers, 2007). Both projects involve corpus annotation: FrameNet is based on the British National Corpus; NomBank uses the Wall Street Journal Corpus of the Penn Treebank. Corpus-based valency lexicons of Slavic languages mostly focus on verbal valency; Polish Valence Dictionary (Walenty[3]) also covers nouns and adjectives (cf. Przepiórkowski & Hajnicz & Patejuk & Woliński & Skwarski & Świdziński, 2014).

Valency of Czech nouns is covered by two valency lexicons that also cover verbs and adjectives, namely a printed dictionary *Slovník slovesných, substantivních a adjektivních vazeb a spojení* (Svozilová & Prouzová & Jirsová, 2005) and an electronic lexicon built during the tectogrammatical annotation of the Prague Dependency Treebank (PDT)[4], called PDT-Vallex[5] (Hajič et al., 2003).

In this paper, we present our work on a new corpus-based valency lexicon of Czech nouns; the lexicographic work on the lexicon started at the beginning of 2016. First, we delimit our theoretical framework (Section 2) and specify typical and special valency behavior of Czech deverbal nouns (Section 3). Then we describe differences between our approach and existing lexical resources for Czech (Section 4, Section 5, Section 6, and Section 7). Finally, we focus on ways of searching for nominal valency patterns through the available Czech corpora, see Section 8.

2. General Framework of Functional Generative Description

Issues of valency of Czech nouns were discussed as early as the 1960s by Jirsová (1966) and Křížková (1968), and the first monograph dealing with valency of non-productively derived Czech nouns was elaborated by Novotný (1980). Valency of nouns is studied within various theoretical frameworks, e.g., the modified valency theory formulated by Karlík (2000), transformational generative grammar (Veselovská, 2001; Dvořáková-Procházková, 2008), the lexicological and "corpus-driven" approach (Čermák, 1991; Čermáková, 2009).

Focusing on deverbal nouns, our approach to noun valency is based on the theory of verbal valency developed within the framework of Functional Generative Description (FGD) by Panevová (1974 and 1975) and Sgall & Hajičová & Panevová (1986). Valency frames are presumably stored in the (mental) lexicon, and are

[1] https://framenet.icsi.berkeley.edu
[2] http://nlp.cs.nyu.edu/meyers/NomBank.html
[3] http://walenty.ipipan.waw.pl/
[4] LDC Catalog No.: LDC2006T01, http://hdl.handle.net/11858/00-097C-0000-0001-B098-5
[5] http://hdl.handle.net/11858/00-097C-0000-0023-4338-F

reflected in the tectogrammatical representation of sentences. Valency frames of verbs contain slots for the following types of complementation:
(a) obligatory and optional inner participants (arguments): Actor (ACT), Patient (PAT), Addressee (ADDR), Effect (EFF), Origin (ORIG);
(b) obligatory free modifications (adjuncts), especially those with the meanings of direction (e.g., *přijet někam*.DIR3 'to arrive somewhere'), location (e.g., *přebývat někde*.LOC 'to dwell somewhere'), and manner (e.g., *chovat se dobře*.MANN 'to behave well').

This distinction between inner participants and free modifications is maintained in the description of nominal valency too. Within the treatment given to nominal valency in the FGD (Panevová, 2000; Kolářová, 2010), the meaning of a given noun is the most important factor in determining its valency frames. Following the definition of syntactic and lexical derivation given by Kuryłowicz (1936), we distinguish three groups of nouns:
(i) Nouns derived from verbs by the so called syntactic derivation, i.e. nouns that differ from their source verbs only in part of speech but not in meaning. Nouns in this group denote actions (*vyrábění / vyrobení* 'manufacturing' // *výroba* 'production') or states (*vyskytování se* / *vyskytnutí se* 'occurring' // *výskyt* 'occurrence'); the forms of their participants are typical (regular).
(ii) Nouns derived by lexical derivation, i.e. nouns whose lexical meaning is unmistakably different from the lexical meaning of their source verbs; it includes names of physical entities related to actions (semantically concrete nouns) such as actor nouns (*učitel* 'teacher'), nouns denoting a thing (*dodávka* 'van', *otvírák* 'opener') or nouns denoting a place (*stoupání* 'slope', *východ* 'exit', *čekárna* 'waiting room'); forms of their participants can be either typical or special.
(iii) Nouns on the boundary between syntactic and lexical derivation. The meaning of these nouns is slightly different from the meaning of action or state nouns, described in group (i); however, the nouns belonging to this category, such as *pochvala* 'praise', are still abstract nouns. The forms of their participants can be either typical or special.

Two basic types of Czech deverbal nouns denote an action or a state and so belong to group (i): nouns derived from verbs by productive means (suffixes *-(e)ní/tí*, as in *honění* 'hunting' or *hubnutí* 'losing weight'); and nouns derived from verbs by non-productive means or by zero suffix (such as *honba* 'hunt', *hon* 'hunt'). These two types of nouns are at the center of attention in this project since they can often exhibit both typical and special valency behavior.

3. Typical and Special Valency Behavior of Czech Nouns

The valency behavior referred to as *typical* can be observed with the nouns derived by syntactic derivation (group (i) in Section 2). When determining their valency frames, the nouns are expected to inherit all participants that are present in the valency frame of their source verbs, including the "verbal" character of the participants such as Actor, Patient, and Addressee. However, the forms of the participants undergo some regular shifts that can be described in terms of rules.

The manifestation of *special* valency behavior is tied with two basic issues (Kolářová, 2010): changes in meaning (e.g., an action → a figurative sense), and characteristic properties of valency complementation. The latter involves three phenomena: special forms of valency complementation (see below), reduction of the number of slots in the valency frame of a noun (either pure reduction or incorporation of a participant), and change of the character of valency complementation to exclusively nominal, as in (2) when compared with (1).

Czech is a highly inflectional language; valency participants of a word are primarily distinguished by their morphological category of case. Following Karlík (2000), a distinction between structural cases (NOM and ACC) and non-structural cases (GEN, DAT, LOC, and INS) is useful for the description of verbal valency. A similar distinction turns out to be important also in the nominal domain. The primary general principle (Karlík, 2000: 184) is as follows: within the process of nominalization, the forms of verbal structural cases change whereas the non-structural cases stay the same. This primary general principle explains the *typical shifts* (e.g., ACC → GEN, *lov velryb* 'a hunt of whales') in the surface forms of participants and makes it possible to describe the valency behavior of most Czech deverbal nouns.

Secondary general principles were formulated by Kolářová (2010); they involve various *special shifts* (e.g., ACC → prepositional phrase, *lov na velryby* 'a hunt for whales, i.e. a whale hunt'). The term "special shift" covers the case in which the form of an adnominal participant differs from the form of the corresponding verbal participant and, at the same time, the new form does not correspond to any of the typical shifts. Special shifts frequently occur with non-productively derived nouns; however, they can also occasionally occur with productively derived nouns.

4. PDT-Vallex

Our approach to the development of a corpus-based valency lexicon extends the approach applied in the PDT-Vallex lexicon (for the differences see Section 5.1). In PDT-Vallex, the core valency information is encoded in valency frames in which the possible alternative forms of complementation are taken into account. PDT-Vallex gives information about semantic roles in the form of tectogrammatical functors of the FGD (Mikulová et al., 2006). Each PDT-Vallex entry describes a lexeme (represented by the "lemma") and its valency frame(s). One valency frame typically corresponds to one meaning (sense) of a word (i.e., a verb, a noun, or an adjective). Although PDT-Vallex does not explicitly work with the term lexical unit, a meaning of a word with its particular valency frame corresponds to a lexical unit, understood roughly as 'a given word in a given sense' (Cruse, 1986).

Concerning nouns, PDT-Vallex 1.0 (included in PDT 2.0) contains 3727 entries. So far, special attention has been paid first to capturing the valency properties of nouns derived from verbs by productive means, such as the noun *balení* 'pack(ing)', and second to nouns occurring as nominal components in support verb constructions such as the noun *nabídka* 'offer' in *učinit nabídku* 'to make an offer'. The delimitation of boundaries between particular meanings of a noun is one of the most difficult tasks in nominal valency description. For example, the noun (lexeme) *balení* 'pack(ing)' is

represented by three valency frames in PDT-Vallex, corresponding to three meanings of the noun, see (1)–(3). Different meanings can sometimes be distinguished by different types or forms of complementation. In (1), we encounter the semantic roles of Actor (ACT) and Patient (PAT), optionally also Effect (EFF); in (2), Material (MAT). The valency frame in (3) is empty.

(1) *balení*$_1$ 'the process of packing':
ACT(GEN,INS,POSS) PAT(GEN,POSS) EFF^{opt}(*na* 'on'+ACC, …)
e.g., *balení dárků*.PAT *rodiči*.ACT 'packing gifts by parents'
(2) *balení*$_2$ 'a container': MAT(GEN)
e.g., *dárkové balení vína*.MAT 'a gift pack of wine'
(3) *balení*$_3$ 'design': an EMPTY valency frame
e.g., *kniha v brožurkovém balení* 'a book in a paperback binding'.

5. The Corpus-based Valency Lexicon of Czech Nouns

In order to create a resource useful to a wide audience (the general public, linguists and applications in second language education and in NLP applications) and to facilitate deeper theoretical understanding of nominal valency, we emphasize the following differences from the existing lexical resources:

5.1 Corpus-based Valency Lexicon vs. PDT-Vallex

Our corpus-based valency lexicon of Czech nouns will give a more elaborate treatment of nominal valency than PDT-Vallex. Although the current version of the nominal entries in PDT-Vallex can be exploited to a large extent, the entries should be improved in several aspects.

Extension of the list of involved nouns. PDT-Vallex covers only nouns that were encountered during the annotation of the treebanks in the Prague Dependency family (PDT, PCEDT, PDTSC)[6]. We plan to treat selected semantic classes more exhaustively, especially if the relevant nouns undergo special valency behavior.

All meanings of a noun. Only the senses which occurred in the annotated data of PDT were included in PDT-Vallex. We plan to provide valency patterns for all meanings of the treated nouns as documented in the much larger Czech National Corpus (CNC)[7] and monolingual dictionaries.

Consistent treatment within semantic classes. PDT-Vallex was built with the intention to enable consistent annotation of each word with its valency in all of the PDT data (Hajič et al., 2003). However, consistency across whole semantic classes went beyond the main goals of the annotation, although it is crucial for the development of theoretical understanding of valency-related phenomena.

Special forms of participants and valency frames. In PDT-Vallex, special forms of participants have been treated mostly as variants of typical forms. However, Kolářová (2014) argues that a participant in a special form cannot co-occur with the same set of forms of other participants as the same participant expressed in a typical form. Such difference in the syntactic behavior of a deverbal noun has a certain impact on its meaning, even if it is only a slight nuance. Consequently, we expect that special forms will be treated in separate valency frames in the valency lexicon.

Participants in combinations. Participants modifying nouns can combine under certain conditions. There are some regular restrictions and rules concerning combinations of particular forms as well as their word order. Especially trivalent nouns constitute rather complex patterns when all three participants are expressed on the surface. We intend to provide all possible combinations of participants in various forms, including word order variants. Further research is necessary in order to determine to what extent can restrictions on combinatorial properties of complementation be treated in a grammar component of the lexicon and to what extent separate valency frames in the data component of the lexicon are necessary (for the two-part model of a dictionary which is divided into a grammar component and a data component, see esp. Lopatková et al., 2015).

Type of special valency behavior. The type of special valency behavior will be specified in the relevant entries (e.g., special form of a participant or reduction in the number of slots).

Frequency or stylistic evaluation of a combination / pattern. Where appropriate, frequency and a stylistic evaluation of a pattern will be indicated. In particular, as we intend to cover all forms encountered in the corpora (if they can be considered grammatical), it is necessary to indicate which of these should be considered central (productive) and which are only peripheral.

Link to the source verb and other deverbal counterparts. Every noun will be provided with a link to the verb the noun is derived from and to other nouns derived from the same verb, especially to both non-productively and productively derived counterparts. As nouns derived by non-productive means are not sensitive to aspect they will be provided by links to both perfective and imperfective verbs.

5.2 Corpus-based Valency Lexicon vs. Slovník slovesných, substantivních a adjektivních vazeb a spojení (SSSAVS)

Our corpus-based valency lexicon will differ from the SSSAVS (Svozilová – Prouzová – Jirsová, 2005) especially in following aspects:

Semantic roles and syntactic ambiguity. Concerning nouns, the SSSAVS represents a traditional way of capturing noun valency by giving only examples of particular complementation, regardless of possible combinations with other types of complementation expressed by various forms, such as *lov na medvěda* 'hunt for bear, i.e. bear hunt', *lov ryb* 'hunt of fish, i.e. fishing' (Svozilová & Prouzová & Jirsová, 2005: 130). The examples convey information about semantic requirements and syntactic forms of the arguments but do not serve as inventory of semantic roles. However, some adnominal forms may occur in constructions that are syntactically ambiguous. This is especially the case of genitives but also of other forms (e.g., in the construction *upozornění řidiče* 'warning of the driver' the genitive form *řidiče* 'of the driver' can be either ACT or ADDR).

[6] http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3, http://hdl.handle.net/11858/00-097C-0000-0015-8DAF-4, http://ufal.mff.cuni.cz/pdtsc1.0/
[7] http://korpus.cz/

Thus we suppose that all valency patterns should be supplemented with the semantic roles of participants.

"Right" as well as "left" valency. The SSSAVS only provides so-called right valency (adnominal counterparts of verbal objects and some adverbial modifications). Adnominal counterparts of verbal agents are not provided at all; the authors suppose that they are used regularly enough and so it would be redundant to supplement all nominal entries with such regular information. However, some combinations of participants are significantly influenced by the presence of an adnominal agent, so we believe that the corpus-based lexicon should provide information on both the "left" and the "right" valency of nouns.

6. Conceptual and Methodological Principles

Vernerová (2011) identifies areas that represent the crucial decisions that have to be made before any lexicographic work is begun. In our lexicon, these decisions will be made as follows:

(a) The lexicon will target linguists as well as non-linguists. We also envisage NLP applications.

(b) The lexicon will be organized into alphabetically ordered lexical entries with additional (optional) semantic classification added to selected lexical units (meanings), similarly as in VALLEX (Lopatková & Žabokrtský & Kettnerová, 2008).

(c) Valency patterns will be described as combinations of syntactic and semantic phenomena.

(d) Valency slots will be identified by semantic roles defined within the theoretical framework of FGD.

(e) The described valency behavior and examples will be based on corpus data, in particular on PDT2.0 and PDT3.0, and on the data from the SYN family of CNC corpora.

(f) The lexicon will provide links to the lemmas of source verbs and to productively or non-productively derived deverbal counterparts.

In addition, we specify the methodology of some partial tasks:

Criteria for the selection of lexemes to be included in the lexicon. In contrast to lexicons that include lexemes exclusively on the basis of their frequency, the lexemes involved in the lexicon will be selected with respect to the following aspects: (i) the semantic class it belongs to; (ii) whether it exhibits typical as well as special valency behavior; (iii) whether it has productively as well as non-productively derived counterparts.

Valency patterns: Utilization of existing valency lexicons. We will look up valency patterns of the source verbs in the existing valency lexicons of Czech, especially in the PDT-Vallex, VALLEX, SSSAVS (Svozilová & Prouzová & Jirsová, 2005) and *Slovesa pro praxi* (Svozilová & Prouzová & Jirsová, 1997). For valency of nouns, we will compare the patterns captured in PDT-Vallex with those present in SSSAVS.

Valency frames. The core valency information will be encoded in valency frames. We will also treat types of valency complementation that are neither participants nor obligatory free modifications but are frequent with the given noun (e.g., *vyznamenání za zásluhy*.CAUS 'award for merits'). This kind of complementation has not yet been treated within PDT-Vallex valency frames but it is incorporated as the so-called "typical" complementation in the valency frames in VALLEX.

Criteria for creating a new valency frame. Kolářová (2014) specifies the following reasons for creating a new valency frame:

(i) a clear change in the meaning of a noun;

(ii) reduction of the number of valency slots;

(iii) a change of the character of valency complementation to exclusively nominal (e.g., Patient → Material);

(iv) a different syntactic behavior of a noun modified by a participant in a special form, when compared with the syntactic behavior of the same noun with the participant in a typical form.

Different meanings of a noun. On the basis of the data of CNC and Czech monolingual dictionaries (especially those available to us in an electronic form), we will identify different meanings of the selected nouns (lexemes). We will focus on lexical units (meanings) with valency potential. We suppose that three basic types of nouns will occur, corresponding to the three types of derivation in Kuryłowicz's sense (Section B.1.1). In particular, the following three prototypical "meanings" are envisaged: (i) an action or a state (the meaning which is parallel to the meaning of the source verb), (ii) an abstract result of an action (an abstract noun), and (iii) a physical object (a concrete noun).

Consistent treatment of a semantic class. We plan to treat especially the following semantic classes in detail: nouns of communication (e.g., *návrh* 'proposal'), psychological nouns (e.g., *obava* 'fear'), nouns of contact (e.g., *dotyk* 'touch') and nouns of exchange (e.g., *výdej* 'distribution'). In order to treat the selected semantic classes consistently we will follow the treatment of verbal valency patterns and their semantic classification applied in VALLEX.

Participants in combinations. The nouns we will focus on are bivalent or trivalent and their participants can often be expressed by several forms that can co-occur in various combinations. However, not all the combinations are grammatical (the systematic restrictions will be specified in the theoretical part of the monograph); for example, this is the case of the combination PAT(POSS) + ACT(GEN) (e.g., **pacientovo*.PAT *ošetření lékaře*.ACT 'patient's treatment of the doctor'), when compared with the combination ACT(POSS) + PAT(GEN) (e.g., *lékařovo*.ACT *ošetření pacienta*.PAT 'doctor's treatment of the patient'). However, there are combinations which are grammatical though not common and their usage should be verified in CNC subcorpora, for example, double post-nominal instrumentals (e.g., *pohrdání názory*.PAT *veřejnosti vládou*.ACT 'contempt for opinions of public by the government'). Combinations that we consider to be grammatical but which do not occur in CNC subcorpora will be labelled by a special mark.

Data format and software. The lexicon will be available as machine readable data in a format suitable for NLP applications.

7. A Nominal Entry Example

Such a detailed and comprehensive description of valency behavior of nouns, fulfilling all the tasks given in Section 6, undoubtedly requires considerable amount of effort. We therefore expect that the lexicon will contain only 400-500 nominal lexemes worked out in full detail,

addressing all of the issues in question, including all meanings of the nouns and a detailed analysis of possible combinations of their participants. We present here the envisaged nominal entry for the noun *vyznamenání* 'honor' which is an example of a noun that is present neither in the current version of PDT-Vallex nor in SSSAVS. This lexeme represents all three types of derivation of nouns (it denotes an action, an abstract result of an action, and also a physical object). It also displays special valency behavior, in particular the special shift in the form of the Patient (ACC → DAT).

Example of a nominal entry:
Noun: *vyznamenání*[pf] 'honor'
Semantic class: evaluation
Source verb: *vyznamenat*[pf] 'to honor'
1. *proces vyznamenání někoho* 'the process of honoring someone'
Frame: ACT(POSS, GEN, INS) PAT(POSS, GEN)
Example: *vyznamenání veterána*.PAT *premiérem*.ACT 'honoring of the veteran by the prime minister'; *vyznamenání premiérem*.ACT 'honoring by the prime minister'; *premiérovo*.ACT *vyznamenání veterána*.PAT 'the prime minister's honoring of the veteran'; *vyznamenání veterána*.PAT 'honoring of the veteran'; ?*premiérovo*.ACT *vyznamenání* 'the prime minister's honoring'; *veteránovo*.PAT *vyznamenání premiérem*.ACT 'the veteran's honoring by the prime minister'; *veteránovo*.PAT *vyznamenání* 'the veteran's honoring';
2. *pocta, vyznamenání udělené někomu* 'honor, award'
Frame: ACT(POSS, GEN) PAT(DAT)
Type of special valency behavior: special form of PAT
Example: *vyznamenání veteránovi*.PAT 'honor addressed to the veteran'; ?*premiérovo*.ACT *vyznamenání veteránovi*.PAT 'the prime minister's honor addressed to the veteran'; ?*vyznamenání premiéra*.ACT *veteránovi*.PAT 'honor of the prime minister addressed to the veteran'.
3. *odznak, medaile, řád* 'badge, medal, order'
Frame: EMPTY
Example: *ověnčený vyznamenáními* 'decked with medals'
4. *nejvyšší stupeň celkového prospěchu* 'honors'
Frame: EMPTY
Example: *studovat s vyznamenáním* 'to study with honors'.

8. Nominal Valency Patterns: Searching through Czech Corpora

To exploit corpus data we will use both methods of searching, i.e. manual searching (Section 8.1) and an automatic preprocessing of corpus evidence (Section 8.2). Examples in the resulting lexicon will be extracted from CNC and/or the Araneum corpus[8] (Benko, 2014).

8.1 Manual Searching

We will take advantage of our experience in searching for nominal valency in lemmatized and morphologically annotated linear corpora such as the SYN family of CNC subcorpora. We will use sophisticated queries that take into account word order variants and include some optional positions (e.g., adjectives modifying the participants) but exclude positions that do not match our requirements (e.g., a verb between the noun and the participant), see the following example of a query searching for adnominal participants in the form of prepositionless genitive:
[lemma="..."] [tag!="[Z|R|V|J].*"]{0,2} [tag="N...2.*"]

However, despite carefully prepared queries, a query can often cover various dependency relations that do not match the intention of the query, so all found occurrences have to be manually checked and evaluated. This method is sufficiently precise but the whole procedure of manual searching is very time-consuming.

8.2 Automatic Preprocessing of Corpus Evidence

We will also exploit the Word Sketches (corpus-based summaries of a word's grammatical and collocational behavior, cf. Kilgarriff and Tugwell, 2001) extracted by Sketch Engine (Kilgarriff et al., 2014). We have[9] access to two large corpora of Czech (and their subsets): the corpus SYN (2.7 Gigawords) and the corpus Araneum Bohemicum Maximum (3.2 Gigawords). However, the Word Sketch Grammars provided for these corpora are not well suited to the analysis of valency behavior of nouns: the grammar provided for SYN does not contain relations for arguments of nouns expressed by some prepositionless cases (dative, accusative, or instrumental, so the types of valency complementation expressed by these cases are completely missing from the Word Sketch), while the grammar provided for Araneum Bohemicum extracts all nouns to the right of the headword within a single relation, listing only their lemmas (not the actual forms), so it obscures the syntactically and semantically crucial distinction between the arguments expressed by different prepositionless and prepositional cases. For example, the WordSketch of the word *dar* 'a gift, a present' lists the lemma *nebesa* 'heavens, paradise' under the relations X Y (immediately following word) and X Nn (a noun within three positions to the right). However, the lemma stands here for three different types of complementation which can be distinguished by their morphological form: genitive case *dar nebes* 'a gift of the heavens', prepositional case *dar z nebes* 'a gift from the heavens', and dative case *dar nebesům* 'a gift to/for the heavens'. For these reasons, we are currently developing extensions of the existing Sketch Grammars which will be more suited to the analysis of noun valency.

9. Conclusion

Our corpus-based valency lexicon of Czech nouns incorporates elaborate and comprehensive theoretical description of valency behavior of Czech deverbal nouns, utilizes existing Czech valency lexicons, and exploits both Czech linear and syntactically annotated corpora. We believe that work on the lexicon will bring new theoretical findings in the field of nominal valency as well as the useful and versatile lexical resource.

[8] http://ucts.uniba.sk/aranea_about/index.html

[9] Because of financial reasons, we depend on the Sketch Engine corpora and functions licensed to the Institute of the Czech National Corpus. Thus, we do not have access to another large corpus of Czech, the czTenTen Web corpus crawled in 2012.

10. Acknowledgements

The research reported in the paper was supported by the Czech Science Foundation under the project GA16-02196S. This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (projects LM2010013 and LM2015071).

11. References

Alexiadou, A., Rathert, M. (Ed.) (2010). *The Syntax of Nominalizations Across Languages and Frameworks*. Berlin: Walter de Gruyter. ISBN 978-3-11-024586-8.

Baldwin, T.; Bond, F. and Hutchinson, B. (1999). A Valency Dictionary Architecture for Machine Translation. In *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, Chester, UK, pp. 207--217.

Benko, V. (2014). Aranea: Yet Another Family of (Comparable) Web Corpora. In P. Sojka et al. (Eds.) *TSD 2014*. LNAI 8655. Springer International Publishing, pp. 247--56.

Chomsky, N. Remarks on Nominalization. (1970). In R. Jacobs, P. Rosenbaum (Ed.) *Readings in English transformational grammar*. Waltham, MA: Blaisdell, pp. 184--221. ISBN 978-0878401871.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge Textbooks in Linguistics. Cambridge, UK: Cambridge University Press. ISBN 0-521-27643-8.

Čermák, F. (1991). Podstata valence z hlediska lexikologického. In D. Rytel-Kuc (Ed.) *Walencja czasownika a problemy leksykografii dwujęzycznej*. Wrocław: Wydawnictwo polskiej akademii nauk, pp. 15--40.

Čermák, F., Holub, J. (2005). *Syntagmatika a paradigmatika českého slova I (Valence a kolokabilita)*. Praha: Karolinum. ISBN 8024609746.

Čermáková, A. (2009). *Valence českých substantiv*. Praha: Lidové noviny. ISBN 978-80-7106-426-800.

Dvořáková-Procházková, V. (2008). Argument structure of Czech event nominals. In F. Marušič, R. Žaucer (Ed.) *Contributions from Formal Description of Slavic Languages 6.5*, Bern: Peter Lang, pp. 73--90.

Dvořák, V. (2014). Case assignment, aspect, and (non-)expression of patients: A study of the internal structure of Czech verbal nouns. In O. Spevak (Ed.) *Noun Valency*, Amsterdam: John Benjamins, pp. 8--112. ISBN 9789027259233.

Grimshaw, J. (1990). *Argument structure*. Cambridge, MA: MIT Press.

Hajič, J. et al. (2003). PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories.* Vaxjo University Press, pp. 57--68. ISBN 91-7636-394-5.

Herbst, T. et al. (2004). *A valency dictionary of English: a corpus-based analysis of the complementation patterns of English verbs, nouns, and adjectives*. Berlin: Walter de Gruyter. ISBN 3-11-017194-5.

Heyvaert, L. (2003). *A cognitive-functional approach to nominalization in English*. Berlin: Walter de Gruyter. ISBN 3-11-017809-5.

Ivanová, M.; Sokolová, M.; Kyseľová, M. and Perovská, V. (2014). *Valenčný slovník slovenských slovies na korpusovom základe*. Prešov: Filozofická fakulta Prešovskej univerzity. ISBN 978-80- 555-1148-1.

Jirsová, A. (1966). Vazby u dějových podstatných jmen označujících duševní projevy. *Naše řeč*, 1966, 49, pp. 73--81.

Karlík, P. (2000). Valence substantiv v modifikované valenční teorii. In Z. Hladká, P. Karlík (Ed.) *Čeština – univerzália a specifika*, 2. Brno: Vydavatelství MU, pp. 181--192. ISBN 80-210-2262-0.

Kilgarriff, A., Tugwell, D. (2001). WORD SKETCH: Extraction and display of significant collocations for lexicography. In *Proc Collocations workshop, ACL 2001*, Toulouse, France, pp. 32--38.

Kilgarriff, A. et al. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), pp. 7--36.

Kingsbury, P.; Palmer, M. and Marcus, M. (2002). Adding Semantic Annotation to the Penn TreeBank. In *Proceedings of the Human Language Technology Conference, HLT-02*. March 24-27, 2002. San Diego, California.

Klímová, J. (2005). Czech lexical database – Derivation Relations. In P. Bouillon, K. Kanzaki (Ed.) *Third International Workshop on Generative Approaches to the Lexicon*, Geneva, May 19-21 2005, Université de Genève, pp. 119--123.

Klímová, J. (2010). Český slovotvorný systém 21. století v databázích. In S. Čmejrková, J. Hoffmannová and E. Havlová (Ed.) *Užívání a prožívání jazyka*, Praha: Karolinum, pp. 147--151. ISBN 978-80-246-1756-5.

Kolářová, V. (2010). *Valence deverbativních substantiv v češtině (na materiálu substantiv s dativní valencí)*. Praha: Karolinum. ISBN 978-80-246-1828-9.

Kolářová, V. (2014). Special valency behavior of Czech deverbal nouns. In O. Spevak (Ed.) *Noun Valency*, Amsterdam: John Benjamins, pp. 19--60. ISBN 9789027259233.

Kuryłowicz, J. (1936). Dérivation lexicale et dérivation syntaxique. *Bulletin de la Société de Linguistique de Paris*. 1936, 37, pp. 79--92.

Křížková, H. (1968). Substantiva s dějovým významem v ruštině a v češtině. In Isačenko, A. V. (Ed.) *Kapitoly ze srovnávací mluvnice ruské a české III. O ruském slovese*, Praha: Academia, pp. 81--152.

Lopatková, M.; Žabokrtský, Z. and Kettnerová, V. (2008). *Valenční slovník českých sloves*. Praha: Karolinum. ISBN 978-80-246-1467-0.

Lopatková, M.; Kettnerová, V.; Bejček, E.; Vernerová, A. and Žabokrtský, Z. (2015). *VALLEX 3.0 - Valenční slovník českých sloves*. Charles University in Prague, [online] http://ufal.mff.cuni.cz/vallex/3.0/.

Melloni, C. (2011). *Event and result nominals: a morpho-semantic approach*. Bern: Peter Lang. ISBN 978-3-0343-0658-4.

Meľčuk, I.A., Zholkovsky, A.K. (1984). *Explanatory Combinatorial Dictionary of Modern Russian*. Vienna: Wiener Slawistischer Almanach.

Meyers, A. (2007). *Annotation Guidelines for NomBank – Noun Argument Structure for PropBank*. [online] http://nlp.cs.nyu.edu/meyers/nombank/nombank-specs-2007.pdf.

Mikulová, M. et al. (2006). Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual. Technical Report TR-2006-30, Praha: ÚFAL MFF UK.

Novotný, J. (1980). *Valence dějových substantiv v češtině*. Sborník pedagogické fakulty v Ústí nad Labem. Praha: SPN.

Osenova, P. (2009). Imennite frazi v bulgarskija ezik. Sofia: ETO.

Panevova, J. (1974 and 1975). On verbal frames in functional generative description. *Prague Bulletin of Mathematical Linguistics*. Part I: 1974, 22, pp. 3--40. Part II: 1975, 23, pp. 17--37.

Panevová, J. (2000). Poznámky k valenci podstatných jmen. In Z. Hladká, P. Karlík (Ed.) *Čeština – univerzália a specifika 2*. Brno: Vydavatelství MU, pp. 173--180. ISBN 80-210-2262-0.

Panevová, J. a kol. (2014). *Mluvnice současné češtiny 2: Syntax češtiny na základě anotovaného korpusu*. Praha: Karolinum. ISBN 9788024624976.

Petkevič, V. (2004). Rule-based Part-of-speech and Morphological Disambiguation of the Czech National Corpus. In *Proceedings of the International Conference "Corpus Linguistics – 2004"*, St. Petersburg: St. Petersburg University Press, pp. 271--285.

Przepiórkowski, A., Hajnicz, E., Patejuk, A., Woliński, M., Skwarski, F. and Świdziński M. (2014). Walenty: Towards a comprehensive valence dictionary of Polish. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis (Eds) *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavík, Iceland: ELRA, pp 2785—2792.

Ruppenhofer, J. et al. (2006). *FrameNet II: Extended theory and practice*. Berkeley, CA: Computer Science Institute, University of California.

Sgall, P.; Hajičová, E. and Panevová, J. (1986). *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. Dordrecht: Reidel. ISBN 90-277-1838-5.

Sommerfeldt, K.E., Schreiber, H. (1977). *Wörterbuch zur Valenz und Distribution der Substantive*. Leipzig: VEB Bibliographisches Institut.

Spevak, O. (Ed). (2014). *Noun valency*. Amsterdam: John Benjamins. ISBN 978-90-272-5923-3.

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

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

Vernerová, A. (2011) Nominal Valency in Lexicons. In J. Šafránková, J. Pavlů (Ed.) *Proceedings of the 20th Annual Conference of Doctoral Students - WDS 2011. Part I: Mathematics and Computer Science*. Praha: MATFYZPRESS, pp. 171--176. ISBN 978-80-7378-184-2.

Vernerová, A.; Kettnerová, V. and Lopatková, M. (2014). To pay or to get paid: Enriching a Valency Lexicon with Diatheses. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*. Reykjavík: ELRA, pp. 2452--2459. ISBN 978-2-9517408-8-4.

Veselovská, L. (2001). K analýze českých deverbálních substantiv. In Z. Hladká, P. Karlík (Ed.) *Čeština – univerzália a specifika 3*. Brno: Vydavatelství MU, pp. 11--27. ISBN 80-2001-2532-8.