Valency Lexicon of Czech Verbs:

Towards Formal Description of Valency and Its Modeling in an Electronic Language Resource

Markéta Lopatková

Habilitation Thesis

Computer Science
Mathematical Linguistics

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Prague 2010

This document has been typeset by the author using LaTeX2e with Geert-Jan M. Kruijff's bookufal class modified by Zdeněk Žabokrtský.
The bibliography has been processed using csplainnat and fullname styles adapted by Zdeněk Žabokrtský.

Abstract

Valency refers to the capacity of verb (or a word belonging to another part of speech) to take a specific number and type of syntactically dependent language units. Valency information is thus related to particular lexemes and as such it is necessary to describe valency characteristics for separate lexemes in the form of lexicon entries. A valency lexicon is indispensable for any complex Natural Language Processing application based on the explicit description of language phenomena. At the same time such lexicons are necessary for building language resources which provide the basis for tools using machine learning techniques.

The present habilitation work consists of a collection of already published scientific papers. It summarizes the results of building a lexical database of Czech verbs. It concentrates on three essential topics. The first of them is the formal representation of valency properties of Czech verbs in the valency lexicon. The logical organization of richly structured lexicon data is presented here. The second topic concerns new theoretical issues that result from the extensive processing of language material, namely the concept of quasi-valency complementation and adequate processing of verb alternations. The third topic addresses questions of formal modeling of a natural language. A new formal model of dependency syntax based on a novel concept of restarting automata is introduced here.

The main applied product of the work presented here is the publicly available *Valency Lexicon of Czech Verbs VALLEX*, a large-scale, high-quality lexicon which contains semantic and valency characteristics for the most frequent Czech verbs. *VALLEX* has been designed with emphasis on both human and machine-readability. Therefore, both linguists and developers of applications within the Natural Language Processing domain can use it.

Contents

Introduction **1**

Selected Aspects of Building the Lexicon **5**
1 What Is Valency? 5
2 Building the First Version of the Lexicon: *VALLEX 1.0* and *1.5* 7
3 Theoretical Aspects of Valency and Their Manifestation in the Lexicon . 15
4 Current Concept of the Valency Lexicon: *VALLEX, Version 2* 21
5 Formal Modeling of Natural Language: Valency as Core Syntactic Information 23
6 Concluding Remarks 29
Bibliography 33

A Building the First Version of the Lexicon **41**
A.1 Enhancing the Valency Dictionary of Czech Verbs: Tectogrammatical Annotation 43
A.2 Valency Dictionary of Czech Verbs: Complex Tectogrammatical Annotation 53
A.3 Valency Lexicon of Czech Verbs VALLEX: Recent Experiments with Frame Disambiguation 63

B Theoretical Aspects of Valency and Their Manifestation in the Lexicon **73**
B.1 Recent Developments in the Theory of Valency in the Light of the Prague Dependency Treebank 75
B.2 Valency Lexicon of Czech Verbs: Alternation-Based Model 87
B.3 Changes in Valency Structure of Verbs: Grammar vs. Lexicon 95

C Current Concept of the Valency Lexicon: *VALLEX, Version 2* **111**
Struktura slovníku *VALLEX* [The Structure of the *VALLEX* lexicon] 111

D Formal Modeling of Natural Language: Valency as Core Syntactic Information **133**
D.1 Modeling Syntax of Free Word-Order Languages: Dependency Analysis by Reduction 135
D.2 Functional Generative Description, Restarting Automata and Analysis by Reduction 145

Remark on Author's Share **165**

Introduction

The language system is usually divided into two basic components – grammar and lexicon. *Grammar* primarily describes general patterns in a natural language in the form of rules which are applicable to whole classes of words (characterized usually by morphological or syntactic features, and sometimes also to a certain extent by their semantic characteristics). *Lexicon*, on the other hand, as it consists of an inventory of language units with their characteristics, usually describes those features of the language system which are tied to particular lexical units. According to the level of description these characteristics can be divided into a group of morphological characteristics which are recorded in morphological dictionaries and a group of characteristics describing the combinatorial potential of separate (lexical) words, i.e., syntactic and syntactico-semantic features; the latter are described in valency dictionaries.

Various theories describing a natural language system differ from each other according to which part of the information needed for the description of the language is captured by general rules (i.e., in the grammatical component of the language system) and which part is best recorded in the form of lexicon entries. Current developments in formal and applied linguistics tend to favor extensive and rich lexicon information (in theoretical studies this tendency is reflected in terms of *lexicalized grammar*, see Joshi, 1985, or *valency syntax*, see Sgall, 1998, 2006).

The habilitation work presented here summarizes the results of building a lexical database of Czech verbs. The topic is highly relevant – at present many linguists, focusing on the description of higher layers of languages, deep (underlying) syntax and semantics, concentrate their attention on the theoretical description of valency (i.e., not only on the valency of central linguistic phenomena, which have been studied since the middle of the last century, but also on valency phenomena on the boundary between the center and periphery of the language as well as on purely peripheral phenomena) and on the description of valency behavior of particular lexemes in dictionaries.

The question of valency and its description is 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).

The project's objective is to create a *Valency Lexicon of Czech Verbs* (henceforth referred to as *VALLEX*, an abbreviation of *VALency LEXicon*), which is described in this collection of papers. *VALLEX* is also the main applied product of the work presented

here. The incentive for the author to propose and work on this project was the absence of a large-scale and high-quality dictionary containing semantic and valency properties of Czech verbs, publicly available and also easily applicable in NLP applications. This project continues the author's large-scale study dealing with one of the sub-problems of syntactic analysis, namely handling prepositional groups (known as 'PP attachment'), which shows how a valence dictionary – together with word order rules, rules using semantic characteristics and rules based on structural word order restrictions – can contribute to the specification of syntactic dependency in prepositional groups. These problems were addressed in the author's PhD. thesis (Lopatková, 2001), which served as the basis for publishing the book *O homonymii předložkových skupin v češtině (Co umí počítač?) [On homonymy of prepositional groups in the Czech language (What can a computer do?)]*, see (Lopatková, 2003b).

The Structure of the Work

The habilitation work submitted here consists of a collection of commented scientific papers which have been already published either as articles or as conference papers.

The introductory text constituting the chapter called **Selected Aspects of Building a Lexicon** starts with an overview of the concept of (verb) valency (Section 1).

Section 2 presents basic strategies used for building the valency lexicon, in particular the conversion of an existing printed dictionary into electronic form and enhancing it with meaning representations – this stage of creating the lexicon is described in detail in Section A.1 which includes the article **Enhancing the Valency Dictionary of Czech Verbs: Tectogrammatical Annotation** (published in the *Proceedings of Text, Speech and Dialogue International Conference, TSD 2001*, see Skoumalová et al., 2001). Automatic extraction of the lexicon from a syntactically or tectogrammatically annotated corpus is an alternative option. Also the relation between the complexity of valency behavior of verbs and the frequency of their occurrence in the corpus is illustrated here. At the end of Section 2 there is a brief description of the first published version of the lexicon, *VALLEX 1.0*. Its main aspects are dealt with in the paper **Valency Dictionary of Czech Verbs: Complex Tectogrammatical Annotation**, here Section A.2 (presented at the International Conference on Language Resources and Evaluation, LREC 2002, see Straňáková-Lopatková – Žabokrtský, 2002). The quantitatively and qualitatively enhanced version of the lexicon, *VALLEX 1.5*, is presented in Section A.3. It consists of the paper **Valency Lexicon of Czech Verbs VALLEX: Recent Experiments with Frame Disambiguation** (published in the *Proceedings of Text, Speech and Dialogue 2005 International Conference, TSD 2005*, see Lopatková et al., 2005a).

Section 3 of the introductory text focuses on theoretical aspects of valency and how they are reflected in the concept of the lexicon. It primarily deals with the original concept of quasi-valency complementation and with problems in distinguishing particular verb senses. Both topics were elaborated in the study **Recent Developments in the Theory of Valency in the Light of the Prague Dependency**

Treebank, here Section B.1 (published in the collection *Insight into Slovak and Czech Corpus Linguistics*, see Lopatková – Panevová, 2006). Other topics presented in this section include the processing of verbs with similar semantic properties and especially the proposal of an alternation-based model of the lexicon. The latter is described in Section B.2, **Valency Lexicon of Czech Verbs: Alternation-Based Model** (the paper was published at the International Conference on Language Resources and Evaluation, LREC 2006, see Lopatková et al., 2006). Classification of alternations has been further studied in relation to *VALLEX* – here it is discussed in Section B.3, **Changes in Valency Structure of Verbs: Grammar vs. Lexicon** (the paper was presented at the Fifth International Conference Slovko held in Smolenice/Bratislava, Slovakia, see (Kettnerová – Lopatková, 2009b).

Section 4 introduces the present version of the lexicon, *VALLEX 2.0*. Its structure is described in detail in the Czech text **Struktura slovníku VALLEX [The Structure of the *VALLEX* Lexicon]**, Chapter C of this work. It is an introduction for a printed version of the lexicon published as *Valenční slovník českých sloves [Valency Dictionary of Czech Verbs]* by Karolinum Press in 2008.

Section 5 deals with formal modeling of a natural language, where valency serves as the essential syntactic information determining the dependency structure of a sentence. The method of analysis by reduction for dependency syntax is presented here and described in detail in **Modeling Syntax of Free Word-Order Languages: Dependency Analysis by Reduction**, here Section D.1 (the paper was published in the *Proceedings of Text, Speech and Dialogue International Conference, TSD 2005*, see Lopatková et al., 2005b). This section also introduces, informally, the concept of a *restarting automaton*, which is able to model the valency syntax and non-local behavior of free-word-order languages in an appropriate way. The reduction system based on the concept of the restarting automaton represents a new formal frame for modeling Functional Generative Description, a theoretical concept of natural language description which forms the underlying theory of this work; the formal model was introduced in the article **Functional Generative Description, Restarting Automata and Analysis by Reduction**, here Section D.2 (presented at the International Conference Formal Description of Slavic Languages 6.5, see Lopatková et al., 2008).

Section 6, concluding the introductory text, summarizes the present use of the *VALLEX* lexicon in NLP applications, mainly in building other electronic lexicons. Directions of further development of the lexicon are also mentioned here, especially gradual and more precise specification of the criteria used for distinguishing particular senses of words and also the influence of the alternation-based model of the lexicon on the lexicon data.

Selected Aspects of Building the Lexicon

1 What Is Valency?

The term valency was first used in chemistry, where it denotes the capacity of an element to combine with a fixed number of atoms of another element. In linguistic contexts it was first used in the middle of the last century by Lucien Tesnière, a French syntactician, as a metaphor for denoting the ability of a verb (similar to the capacity of atoms) to bind a certain number of language elements.

The concept of valency refers to "the range of syntactic elements either required or specifically permitted by a verb or other lexical unit ..." *Concise Oxford Dictionary of Linguistics*, see Matthews, 1997). It is the ability of a word to open a certain number of positions for other, syntactically dependent language units. In our approach, this ability concerns primarily the underlying layer of language, i.e., the layer of deep syntactic structures.[1] Valency positions are occupied by valency complementations such as Actor (Agent or Bearer of an action, denoted as ACT here), Patient (Object affected, denoted as PAT), Addressee (ADDR), Origin (ORIG) and Effect of an action (EFF) that are usually denoted as inner participants (also actants, arguments) and by free modifications expressing the circumstances such as time, place, direction, manner etc.

The set of valency positions characteristic for a word in one of its senses is called a *valency frame* of a given word in a given sense.[2]

Particular valency positions are characterized by different levels of semantic obligatoriness. If a certain position remains unfilled, the semantic completeness may be distorted; this may result in a grammatically incorrect sentence, cf. unacceptable sentences such as **Petr dává [Peter gives]*, **Marie nenávidí [Mary hates]*, or **Jan se choval [John behaved]*. Such complements are called obligatory (at the layer of deep syntax). Particular attention is paid to cases when the obligatory valency positions remain unoccupied in the surface form of the sentence: the participant is generalized,

[1] Different 'levels' of valency are sometimes distinguished (Matthews, 1997): *syntactic valency* (also grammatical valency) working with concepts like subject and object and *semantic valency* (also lexical valency) working with terms like semantic roles or case roles as, e.g., Agent, Experiencer, Patient or Goal (similar to semantic cases, thematic roles, theta roles (generative grammar) or deep case (case grammar), depending on the background theory). Compare also distinction between *valency* (Cz. 'valence') and *intention* (Cz. 'intence') usual in Czech linguistics (Daneš et al., 1987).

[2] The relation between separate senses of a word and its valency frames is generally more complex, see Section 3 (subsection dealing with meaning specification and distinguishing separate verb senses) and Chapter C (Remark on page 116-117).

as e.g. in *Neruš Petra, čte (= něco) [Don't disturb Peter, he is reading (= something)]*, or the receiver fills the positions on the basis of the discourse context, e.g., in sentence *Děti přišly [Children came]* the obligatory direction complementation is omitted – this information is assumed to be deducible from the context of the discourse.

There are other valency positions that are optional – they are present in a valency frame and they may be present in the meaning representation of a sentence; however, their omission does not result in semantically or grammatically incorrect sentences, e.g., *Petr se pevně držel (zábradlí) [Peter held on firmly (to the railing)]*, *Eva se najedla (ovoce) [Eve has eaten (some fruit)]*, *Dívka píše (mamince) dopis [A girl is writing a letter (to her mum)]*. Yet another set of positions is characterized by a very loose relation to the verb; these are usually referred to as non-valency or free modifications, e.g., *Jana se procházela (po lese) [Jane walked (in the wood)]*, *Petr se budil (časně) [Peter woke up (early)]*, *Eva si četla (pro své potěšení) [Eve was reading (for pleasure)]*, although, in the theoretical description, they are included and treated within verb valency (in a broad sense).

In the surface realization of a sentence particular participants usually occur in a certain form which is affected by the verb – their morphemic form is determined by the requirements of the governing verb. Thus, the Actor in an active sentence, for example, is prototypically expressed by the nominative while the Patient is usually realized as the accusative, e.g., *Petr.*ACT-nom *ztratil botu.*PAT-acc *[Peter.*ACT *lost his shoe.*PAT*]*. Other verbs require the Actor in the dative, e.g., *Petrovi.*ACT-dat *se ve škole líbí [Peter.*ACT *likes the school]*, and still other verbs require the Patient to be in the dative, e.g., *Rodiče.*ACT-nom *bránili jejich štěstí.*PAT-dat *[Parents.*ACT *obstructed their happiness.*PAT*]*, or in the form of a prepositional group, e.g., *Doufali ve vítězství.*PAT-v+acc *[They hoped for victory.*PAT*]*. On the other hand, forms of free valency modifications are usually determined by their meaning, e.g., *Děti přišly domů / do školy / na hřiště [Children came home/to school/to the playground]*, and not by the grammatical properties of the governing verb.

Valency has been studied theoretically from the middle of 20th century; here the names of L. Tesnière (1959) and and M.A.K. Halliday (1963) should be mentioned. The semantic aspects of valency were discussed especially by Ch. Fillmore (1968; 1969). In Czech linguistics, we can mention mainly F. Daneš, e.g., (1971; 1987), J. Panevová, esp. (1974-5; 1980; 1994; 2000), P. Karlík (2000), a and P. Sgall (1998; 2006). Further references can be found in Section C – an introductory text for the printed edition of the dictionary *Valenční slovník českých sloves [Valency Dictionary of Czech Verbs)]*.

A verb is traditionally considered to be the structural center of a sentence because the relational structure within a sentence is formed by the verb's valency requirements, i.e., its requirements for the number and the type of syntactically dependent language units (based on *Mluvnice češtiny 3*, Daneš et al., 1987). That is the reason why valency theory focuses primarily on verbs; the valency of other types of lexical words (autosemantic parts of speech, i.e., nouns, adjectives and adverbs) is usually considered to be secondary.

It is obvious that valency properties of words are considerably diverse. They cannot be derived by means of general rules; it is necessary to describe them separately for each individual lexical item, i.e., in the form of a valency lexicon which describes the valency of words one after another, in each of their senses. The verbocentric approach in theoretical syntax is also reflected in the efforts to build valency lexicons primarily for verbs; for the description of the valency of deverbal nouns and adjectives[3] the structural similarity between these units and verbs is used most frequently (see also Section 6).

It is necessary to mention that nowadays the building of valency lexicons in both printed and electronic forms is in the center of attention of Czech lexicography as well as the lexicography of many other languages; references to the most significant lexicons are provided in Section C, an outline and summaries of the most prominent projects dealing with valency can be found in (Lopatková et al., 2002b) and (Žabokrtský, 2005).

2 Building the First Version of the Lexicon: *VALLEX 1.0* and *1.5*

The development of computational linguistics and interest in applied tasks have led to a need for linguistic resources containing various types of linguistic information (morphology, surface or deep syntax, disambiguation etc.). Especially thanks to the Prague Dependency Treebank (PDT, Prague Dependency Treebank, see Hajič, 2006), the Czech language is one of the languages with the most extensive data resources.

A valency lexicon represents a significant language resource which provides important information for many NLP tasks. Valency plays a key role in complex applications working with meaning, such as automated translation or summarizing, where syntactic analysis and automatic disambiguation of particular meanings of lexical words belong to the essential prerequisites.

Conversion of the Printed Dictionary and Its Enhancement by Deep-Syntactic Representation

The basic method for obtaining a valency lexicon is conversion of a printed normative dictionary into electronic form followed by the extraction of valency information. This method was used with the *BRIEF* lexicon (Pala – Ševeček, 1997), which is based mainly on *Slovník spisovného jazyka českého*, denoted here as SSJČ (Havránek, 1964). The text data were processed automatically and possible surface combinations of valency complementations were extracted from the dictionary entries (however, separate senses of the verbs were merged in the process). The lexicon obtained in this way is limited to the description of surface valency information (mainly in the form of morphological cases and prepositional groups) and does not deal with the obligatoriness and optionality of separate valency complementations. The main drawback of the *BRIEF* lexicon is the loss of the information about individual verb senses. On the other hand, formal representation of frames, described e.g. in (Horák, 1998), makes it possible to use the lexicon in NLP applications.

[3] Here we completely leave aside other types of nouns and adjectives, as well as adverbs.

Automatic enhancing of the *BRIEF* lexicon with the missing valency information appeared to be a natural solution. Such automatic enhancing of the lexicon was attempted by H. Skoumalová (2001). She focused on the automatic delimitation of senses and addition of meaning characteristics in the form of functors for each valency position; the basic algorithms were introduced in (Panevová – Skoumalová, 1992). Later a test set of verbs, denoted as *Vallex-00*, was processed with the use of these algorithms. The test set consisted of the 178 most frequent Czech verbs (excluding the verb *být [to be]* and modal verbs) and was processed first automatically and then refined manually on a large scale. This stage is described in more detail in Section A.1. It was shown that the necessary manual editing[4] was so extensive that it is more efficient and more convenient for the annotators to build the lexicon manually from scratch.

Valency Information and the Syntactically Annotated Corpus

Another option for building a valency lexicon is to use existing syntactically annotated corpora. The computational linguistics literature provides descriptions of several methods for automatic extraction of at least surface valency information ('subcategorization frames') from annotated corpora. These methods were tested also for the Czech language. What should be mentioned here are particularly the attempts described in (Zeman – Sarkar, 2000; Sarkar – Zeman, 2000), which report on constructing surface valency frames from the analytical layer of PDT (Hajič, 1998).

Frames obtained in this way do not contain the meaning characteristics of separate valency positions and do not tackle their obligatory nature. Again, the main defect here lies in the fact that the frames obtained do not correspond to separate verb senses.

Valency Information and the Tectogrammatically Annotated Corpus

The third option for building a valency lexicon is direct extraction from a tectogrammatically annotated corpus, i.e., a corpus enriched with deep-syntactic information describing the linguistic meaning of sentences (according to Sgall et al., 1986b). The essential idea is simple – if we have a corpus annotated by experienced annotators-linguists at the tectogrammatical layer, we have also the information about each verb's valency characteristics. That is to say, we get the information about the number and the type of valency positions and, secondarily, also about their obligatoriness (the same holds true for nouns or adjectives with valency requirements). Then one can simply collect this information in the form of a valency lexicon.

However, in the course of the extensive manual annotation of the tectogrammatical layer of PDT (Sgall et al., 2004; Hajič, 2006) it was found that the annotators, although knowledgeable linguists, need a valency lexicon for their work because in the course of the complex tectogrammatical annotation they are not able to concentrate on the

[4] This concerns especially the changes related to the delimitation of verb senses - approximately 350 automatically proposed frames were extended by the annotators to more than 460 frames, which roughly correspond to separate verb senses.

problems of separate linguistic phenomena. Thus it was decided to build the valency lexicon manually with as much technical support as possible (appropriate graphical annotation interface, a variety of electronic resources, searching according to various criteria, check for consistency etc.).

The first stage was represented by collecting valency frames from the above-mentioned set of verbs, *Vallex-00*, see Section A.1, and also from the additional lists used by the annotators while annotating PDT (December 2001). The material so obtained was processed thoroughly. First, the frames corresponding to each other were identified and additional criteria for sense disambiguation were stated. After being processed in this way the verbs (331 in total) became the core for building an electronic valency lexicon.

Considering the necessity of using the valency lexicon for the annotation of PDT, its construction proceeded along two lines. Both these lines, including the theoretical background, are explained in (Lopatková, 2003a), so only a brief description of them is provided here.

PDT-VALLEX. The *PDT-VALLEX* lexicon describes valency frames of the verbs which occurred in the course of the annotation of PDT, but only in those senses in which the relevant verbs occurred in PDT, with certain exceptions.[5] Individual occurrences of verbs in PDT are linked with the lexicon entry. *PDT-VALLEX* was constructed gradually as an annotation tool ensuring the data consistency and after finishing the annotations it underwent an extensive check for consistency; the modifications/corrections were projected back into the PDT data. This line of the lexicon is not treated here in more detail, for further information see esp. (Hajič et al., 2003; Urešová, 2006, 2009).

VALLEX. The *VALLEX* lexicon aims at describing valency behavior of verbs in each of their senses,[6] i.e., at providing analysis of whole verb lexemes. In addition to valency frames, further syntactic information is rendered for each sense of a verb, i.e., the information related to the surface manifestation of verb valency (e.g., reciprocity, reflexivity, grammatical control), and also some syntactico-semantic information (primarily a syntactico-semantic class for a substantial subset of verbs). Adequate theoretical description of valency characteristics, relative comprehensiveness and consistency of processing are stressed in *VALLEX*.

The current form and size of *VALLEX* lexicon are described in detail in Section C of this work.

At present a project is in progress, the goal of which is to interlink (semi-)automatically both lines of the lexicon, *PDT-VALLEX* and *VALLEX* (Bejček, 2009). The result of this project will be a valuable new linguistic resource describing

[5] *PDT-VALLEX* contains also valency frames of certain types of deverbal nouns and adjectives.

[6] *VALLEX* concentrates on both primary and secondary meanings; the description of idioms and phrasemes is not covered completely.

the valency characteristics of verbs (predominantly), which will be widely interlinked with corpus annotations.

It is necessary to say that at the same time and facing the same problems, a valency lexicon for the English language, *PropBank Lexicon* (Kipper et al., 2004) is being constructed. This lexicon is connected with the *Proposition Bank corpus* (PropBank, see Palmer et al., 2005) and is based on the annotation of (so-called) propositions and their argument structure in the *Penn Treebank* (Marcus et al., 1993).[7]

Coverage of Verb Occurrences in a Corpus and the Complexity of Valency Behavior of Verbs

Although the manual building of an extensive valency lexicon is time-consuming, it guarantees the required consistency and adequacy of verb description. Analysis of Czech texts shows that the coverage of texts for the most frequent Czech verbs tends to follow Zipf's law (Zipf, 1935), i.e., separate verbs occur in a certain statistical distribution – roughly speaking, the frequency of a verb in the corpus is inversely proportional to its rank in the frequency table, see Figure 1.

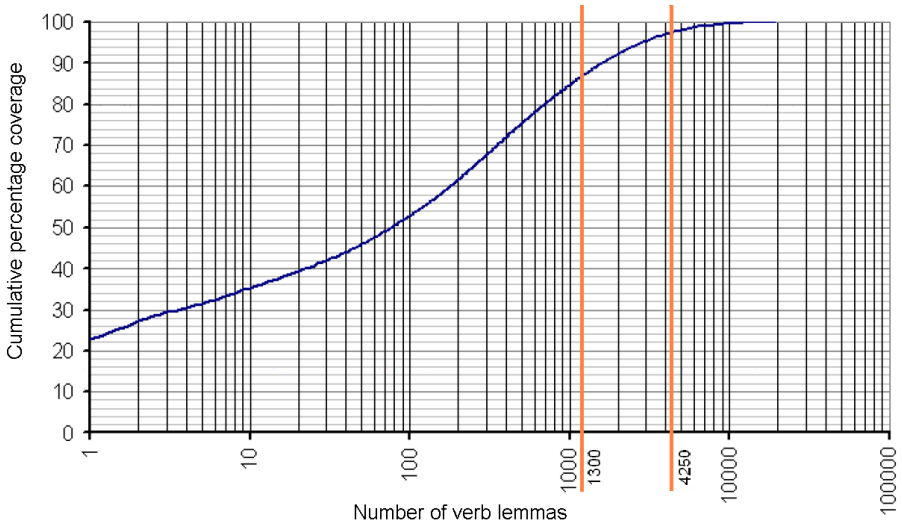


Figure 1: The coverage of the sub-corpus of ČNK by verb lemmas (chart adopted from Žabokrtský, 2005).

The chart shows the number of verb lemmas (without reflexive morphemes *se/si*) on the horizontal axis (with logarithmic scale) and the cumulative percentage coverage of the corpus[8] on the vertical axis. Looking at the chart, it can be seen that, for example, a valency lexicon

[7] In the *PropBank Lexicon*, each verb is represented by a frame composed of one or more 'framesets', which refer to individual verb senses. Each frameset consists of a set of semantic roles ('rolesets') for arguments labeled Arg0, Arg1, ..., ArgM and a number of functional tags for adjunct-like modifiers. *PropBank lexicon* relies rather on syntactic than semantic criteria, which results in coarse-grained rolesets (see also Section 6).

[8] In this orientative chart all occurrences of verbs are counted, including those with auxiliary verbs.

containing the 1300 most frequent Czech verbs covers approximately 85% of the occurrences of verbs in the sub-corpus of the *Czech National Corpus* (ČNK, SYN2000).[9] A valency lexicon of this size roughly corresponds to *VALLEX 1.0* extended with the verb *být [to be]* and modal verbs (the versions of the dictionary are described below). The valency lexicon describing 4250 Czech verbs covers more than 96% of occurrences of verbs in this sub-corpus (a lexicon of this size corresponds to *VALLEX, version 2*; see Section 4).

If we continue examining the complexity of valency behavior of verbs, we can find another significant general characteristic: the higher the frequency of a verb in the corpus, the more valency frames it has and the more complex its valency behavior is;[10] see the chart in Figure 2.

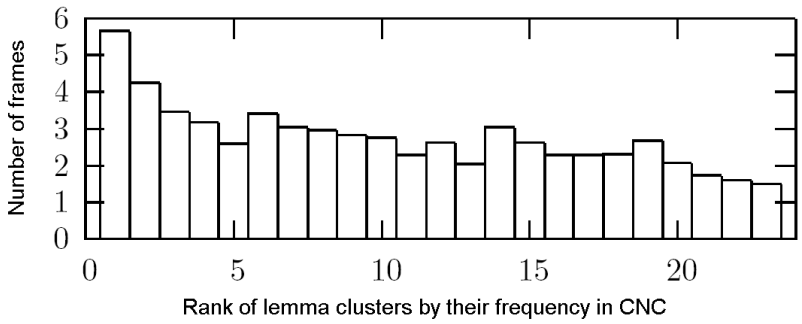


Figure 2: The number of valency frames for the lexemes in *VALLEX 1.0* and their frequency in ČNK (adopted from Bojar et al., 2005).
Separate lexemes (in which both aspects of a verb, i.e., perfective and imperfective, see Section 4 below, are grouped together; they are denoted as 'lemma clusters' in the graph in Figure 2)[11] were ranked in decreasing order of frequency and divided into groups of 40 lexemes. The chart bars show the average number of valency frames for these groups of lexemes ('number of frames').[12]

There is yet another characteristic of verb valency, which is the distribution of valency frames in relation to separate lexical units (LUs, see Section 4) among separate verbal lexemes. The charts in Figure 3 how the inversely proportional relation between the number of lexemes and the complexity of their valency behavior - most verbs have

[9] `http://ucnk.ff.cuni.cz/`

[10] Naturally, this characteristic cannot be mechanically applied to individual verbs, one should read it as a general trend, which is revealed in statistical comparison of lexicon data.

[11] Although verbs forming aspectual pairs are captured separately in *VALLEX, version 1*, it is advisable to treat them together as one lexeme, see especially Section 4. When selecting verbs for the processing, the aspectual counterpart for each highly frequented verb was added even if its frequency was lower. In spite of a lower frequency, these verbs exhibit greater complexity, which is caused by their syntactical similarity to their more frequent aspectual counterparts.

[12] The number of valency frames for a given lexeme is counted as the sum of valency frames for separate aspects.

one valency frame (i.e., they are so-called monosemic lexemes), quite a lot of verbs have only a few valency frames; and, on the other hand, a small number of verbs prove to be considerably complex in their valency behavior.

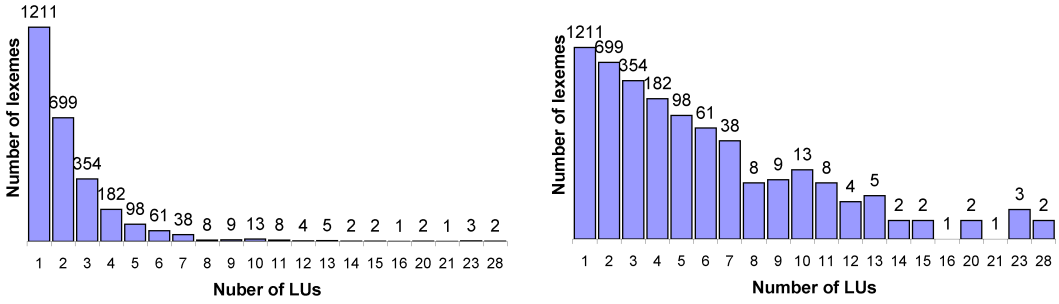


Figure 3: The number of lexemes in *VALLEX, version 2* in relation to the number of lexical units (number of LUs in linear scale on the left; the number of LUs in logarithmic scale on the right). The charts are based on the number of LUs in *VALLEX 2.5*, where aspectual counterparts are associated in one lexeme. Thus, if two or more verbs, being the aspectual counterparts, share the same LU, then it is counted as one LU.

This observation implies that the demanding manual preparation of the lexicon is well founded up to a point. The most frequent verbs show the most complex valency behavior and the extraction of their characteristics by the automatic processing of the existing data is not satisfactory. Consequently, it is reasonable to process them manually and get reliable and consistent descriptions for them. Moreover, the coverage of verbs in texts grows rapidly. The decreasing frequency of a verb in the corpus corresponds to the gradual decreasing of its (average) complexity; this means that very rare verbs may be expected to be characterized by simple valency behavior that can be obtained by (semi-)automatic methods. However, there is no apparent boundary between verbs with complex valency behavior and 'simple' verbs. The graph in Figure 2 shows that the verbs of the last group, i.e., the verbs ranked at about 1000 in ČNK, have about 1,5 valency frame in average. In spite of this some very complex verbs occur in this group, e.g., *vytáhnout* ^{pf} *[to pull out]* (position 1000 in ČNK, in descending frequency order) has 13 valency frames for its non-reflexive lemma and a further 3 for its reflexive lemma, 9 of which are used idiomatically; this verb shares 10 of the total number of its frames with its imperfective counterpart *vytahovat* ^{impf} (position 1803 in ČNK); in this case a large number of the valency frames representing individual lexical units results from the ambiguous prefix *vy-* and from quite a lot of idiomatic frames (the numbers are based on *VALLEX 2.5*).

The First Released Version of the Lexicon: *VALLEX, Version 1*

Intensive work on the manual building of *VALLEX* lexicon started in 2001. Firstly, a detailed report was elaborated (Lopatková et al., 2002b) summarizing different ap-

proaches to valency description around the world (for English, German, Polish, Slovak, Russian, Bulgarian and Japanese) and in the Czech Republic (above all, the theory of sentence patterns, see Daneš – Hlavsa, 1987, valency theory in Functional Generative Description, see Panevová, 1974-5, 1980, 1994). This report provides a description of particular phenomena captured in the lexicon (especially valency frames extended with quasi-valency and typical complementations and other phenomena related to valency such as reflexivity, reciprocity, control and aspectual pairs). It also introduces the tools developed to help the annotators, mainly a text editor with syntax highlighting, a www interface for searching in electronic language resources (especially in the dictionaries such as *SSJČ*, see Havránek, 1964, *BRIEF*, see Pala – Ševeček, 1997, *Slovesa pro praxi*, see Svozilová et al., 1997, and in the sample of Czech National Corpus) and, last but not least, also the advanced searching interface for the *VALLEX* data. The above mentioned facilities together with the XML data structure of the lexicon are described in detail in (Žabokrtský, 2005).

The first released product of the project (and the first machine-readable lexicon with valency representation of Czech verbs) was *VALLEX, version 1.0*, which was made accessible on the website of the Institute of Formal and Applied Linguistics at the Faculty of Mathematics and Physics, Charles University in Prague (MFF UK) in 2003.[13]

VALLEX 1.0 contained the description of valency characteristics for approximately 1400 Czech verbs – the core, *Vallex-00*, was extended and the 1000 most frequent verbs and their aspectual counterparts (according to their frequency in ČNK, with the exclusion of the verb *být [to be]*) were added. There are nearly 2500 valency frames related to these verbs. The verbs included in *VALLEX 1.0* cover approximately 56.4% of verb occurrences in the mentioned sub-corpus of ČNK (further 28.5% relate to the verb *být [to be]* and modal verbs), see the chart in Figure 1.

The structure of *VALLEX 1.0* is described in detail in (Žabokrtský – Lopatková, 2004) and in the Help page for HTML format of the lexicon, as outlined in Section A.2 here. The topmost level of the lexicon is formed by verb entries represented by separate verb lemmas (morphological variants of a lemma are described within one entry). Each verb entry consists of a non-empty sequence of valency frames ('frame entry', see Figure 4, showing typically one of the senses of a verb). In addition to the valency frame itself, the entries include also obligatory attributes (glosses, examples) and optional attributes (especially control, syntactico-semantic class and reference to the aspectual counterpart).

From the very beginning, *VALLEX* was designed with an emphasis on both human and machine-readability. For this reason *VALLEX 1.0* was released in three formats:

HTML format. *VALLEX 1.0* in the form of a web application, enabling convenient search in the lexicon according to various aspects (besides the verbs in alphabetical order it is possible to search also according to functors,[14] forms of comple-

[13] `http://ufal.mff.cuni.cz/vallex/1.0/`

[14] Functors indicate the types of syntactico-semantic relations between a verb and its complementa-

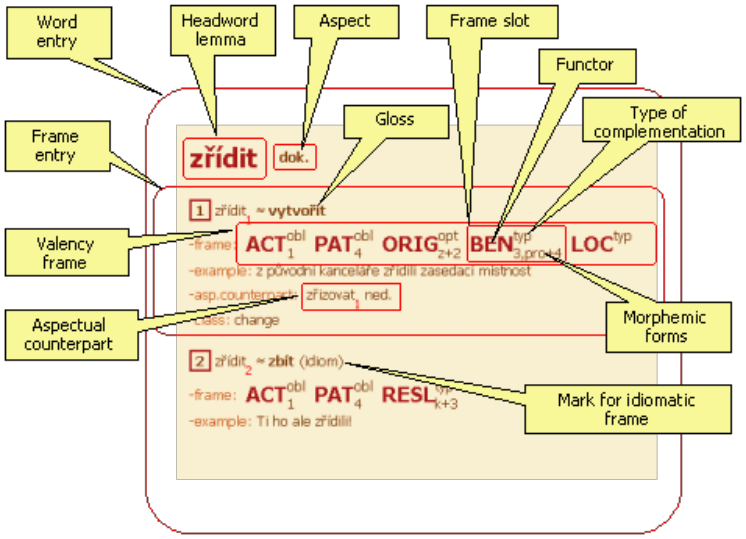


Figure 4: The structure of a lexicon entry in *VALLEX, version 1* (adopted from the Help page for HTML format of the lexicon).

mentation, syntactico-semantic classes or control).

PDF format. *VALLEX 1.0* in a printable version (published also as a technical report, see Lopatková et al., 2003).

XML format. XML is the primary format of *VALLEX 1.0*, designed for computer applications and advanced search.

Processing a large amount of data proved that the conception of the lexicon developed here is productive and that the methodology for processing verb entries meets the requirements for effective and consistent lexicographical work (the process of testing the consistency and completeness of *VALLEX 1.0* was described in Lopatková – Žabokrtský, 2003). All this was reflected in the users' interest in *VALLEX 1.0* and made it possible to proceed with the quantitative and qualitative extension of the lexicon. The first result of this was a new working version, *VALLEX 1.5*.

VALLEX 1.5 consisted of 2500 verbs (approximately 6000 valency frames), including the verb *být [to be]* and modal verbs. This version was made accessible on a restricted-access basis to the users at the Institute of Formal and Applied Linguistics in Prague and was used for extensive testing of the selected formats. Great emphasis was put especially on the consistent and systematic capturing of the various phenomena; this version provided valuable feedback revealing problematic former decisions and non-systematic solutions and enabled further quality improvement. *VALLEX 1.5* is described in Section A.3. Also the first results of automatic frame disambiguation, i.e.,

tion, e.g. ACT for Actor, PAT for Patient, DIR3 for the complementation of Direction-where to; see the list of functors in Chapter C, pp. 120-121.

matching of valency frames to particular verb occurrences in a text are presented there (for more details see Concluding Remarks in Section 6).

3 Theoretical Aspects of Valency and Their Manifestation in the Lexicon

The theoretical basis for building the *VALLEX* lexicon was provided by Functional Generative Description (FGD), see especially (Sgall, 1967; Sgall et al., 1986b); FGD is characterized by its dependency and stratification-based approach to language description. Valency theory as one of the core concepts of FGD has been developing since 1970s; a substantial summary is provided in (Panevová, 1994). This conception represents a rather syntactically oriented approach to valency. Valency complementations are sorted into *inner participants* and *free (adverbial) modifications* on the basis of their syntactic behavior, see below. Five inner participants have been determined – Actor (labeled ACT, corresponding prototypically to the first syntactic position, i.e. Subject position), Patient (PAT, prototypically direct Object) and Effect (EFF) are semantically indistinctive; on the other hand, Addressee (ADDR) and Origin (ORIG) have typical semantics. Similarly, free modifications are semantically homogenous. Furthermore, obligatory (in deep representation) and optional complementations are distinguished. *Valency frame* is then defined as a set of inner participants, both obligatory and optional, together with obligatory free modifications.

In the course of building the *VALLEX* lexicon, the theoretical framework of valency conception developed within FGD was applied to a large amount of data. Verb lexemes are processed here in their complexity, while in theoretical works mainly primary meanings of verbs were studied and described. These issues require more precise specification of functional criteria defined by theoretical research as well as defining of further criteria.

The fundamental theoretical aspects of building a valency lexicon are described in (Lopatková, 2003a). It provides a basic outline of the concept of quasi-valency complementation and attention is paid also to the delimitation of separate verb senses. Both of these topics were further studied in greater detail on a large amount of corpus data – a detailed description is provided in Section B.1 of the presented work.

Quasi-Valency Complementations

The term quasi-valency was introduced in (Lopatková, 2001) and subsequently in (Lopatková, 2003b). It extends the existing concept of valency as a set of inner participants and obligatory free modifications to include complementations used frequently with a given verb. A large amount of data, processed in the course of building the lexicon and also during the PDT annotation, showed that there is a group of complementations which do not meet the strict criteria for inner participants but are still lexically bound. Their properties are similar to inner participants (ACT, PAT and EFF in particular):

- their morphemic form is determined by the requirements of a governing verb;
- they modify a limited (more or less closed) class of verbs;
- as a complementation of a given verb, they cannot be repeated (except for the case of coordination).

On the other hand, their other characteristics make them similar to free modification (see also Chapter C, where the criteria distinguishing inner participants and free modifications are formulated):

- they are semantically homogenous;
- they are mostly optional;
- they do not undergo the 'shifting' (see Panevová, 1974-5).

The following types of syntactico-semantic relations were classified as quasi-valency complementations: a complementation of intention for certain verbs of movement (INTT, e.g. *Petr mamince doběhl nakoupit [Peter went shopping]*), a complementation of obstacle for the sub-class of contact verbs (OBST, e.g. *Chlapec zakopl o kořen [The boy stumbled over a root]*) and a complementation of difference for the verbs expressing change of state (DIFF, e.g. *Hodnota akcií stoupla o 100 % [The share value increased by 100%]*); also a complementation of mediator (MDT, e.g. *Když jsem odcházel, zatahal mě soused za rukáv [When I was leaving my neighbour tugged at my sleeve]*) (this type of complementation has not been applied to the lexicon data yet).

A question how to classify Addressee (ADDR) and Origin (ORIG) was also opened in this article. Both of these complementations are ranked among inner participants in 'classical' FGD theory; however, their semantic homogeneity indicates that they share also the most important characteristic of quasi-valency complementations.

The newly proposed concept of quasi-valency complementation enriches the traditional division of verb complementations and facilitates more subtle classification and appropriate description of complementations on the boundary between inner participants and free modifications.

Meaning Specification and Distinguishing Separate Verb Senses

Valency frames (sets of valency positions characterized by functors representing the syntactico-semantic relation of the complementations to the verb, by possible morphemic forms and level of obligatoriness) are understood as elementary syntactico-semantic information characterizing separate lexical units. In the *VALLEX* lexicon conception, valency frames roughly correspond to separate senses of a given verb.[15]

There are no generally accepted testable functional criteria for distinguishing individual senses of a given verb; the transition from one sense to another is gradual,

[15] The question of the relation between valency frames and separate senses of a verb is more complex, see Chapter C dealing with current structure of the dictionary (namely Remark on Delimitation of Separate Senses on pages 116-117), and especially the concept of alternations presented below.

without strict boundaries between separate senses in many cases. In the *VALLEX* lexicon, emphasis is put on syntactic criteria, especially the structure of a valency frame, when distinguishing separate verb senses. At the same time the semantics of a verb is taken into account. Two fundamental principles for delimiting separate verb senses are defined in Section B.1:

- Any change in a valency frame, i.e., any change in number and type of complementations (with the exception of possible morphemic variants in the realization of individual functors) results in the specification of a new valency frame (see Section B.2), and thus a new LU.
- Any significant change in meaning of a verb inevitably requires the specification of a new LU (and thus also a new valency frame of the verb even if its syntactic structure remains the same).

The latter principle enhances significantly the earlier rather syntactic approach to valency where the LUs described by the same valency frame remain indistinct despite being semantically different. Such a modification facilitates to reflect not only syntactic but also semantic features of verbal complementations. At this point it is apposite to mention at least the most significant approaches providing such type of information, namely *FrameNet*[16] (Fillmore et al., 2003; Ruppenhofer et al., 2006) and *Pattern Dictionary of English Verbs* (Hanks – Pustejovsky, 2005) based on *Corpus Pattern Analysis*[17] Hanks (2004, 2010), and to a certain extent also *WordNet*[18] (Fellbaum, 1998; for Czech WordNet Pala – Smrž, 2004), see also Section 6.

Processing of Verbs with Similar Semantic Properties

One of the topics which have been widely discussed in recent times is the relation of syntactic and semantic properties of verbs, see especially (Levin, 1993) and (Levin –

[16] *FrameNet* (`http://framenet.icsi.berkeley.edu/is`) is an on-line lexical resource for English, 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 SF is conceived as a "conceptual structure describing a particular type of situation, object, or event" (Ruppenhofer et al., 2006). Each SF contains the so-called frame elements (FEs), i.e., semantic participants of such situations.

[17] According to (Hanks – Pustejovsky, 2005), *Corpus Pattern Analysis* (`http://nlp.fi.muni.cz/projects/cpa/`) 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. 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).

[18] *WordNet* `http://wordnet.princeton.edu/` is a large lexical database of English that groups nouns, verbs, adjectives and adverbs into "sets of cognitive synonyms (synsets), each expressing a distinct concept".

Hovav, 2005). In the course of processing lexicon entries, the examination of the valency of whole groups of verbs with similar semantic properties proved to be effective. Although none of the existing classifications of verbs[19] can be easily adopted for this purpose, about twenty relatively rough syntactico-semantic groups were proposed in *VALLEX* that group together verbs with similar or identical syntactic behavior. These groups serve as a helpful starting point for more detailed examination of syntactic and semantic verb properties; e.g., analysis of some of the verbs of exchange in (Lopatková – Panevová, 2004) or the analysis of selected prefixed verbs of movement in (Lopatková – Panevová, 2007). Another interesting group of verbs being studied in relation to the *VALLEX* lexicon is a large group of verbs of communication, which are characterized by a propositional complementation (so-called 'verba dicendi' in Slavic linguistics, compare also 'speech act verbs' in Wierzbicka, 1989). For example, (Kettnerová, 2008) propose the specification of several subclasses of verbs of communivation based on the sentence modality of a dependent clause realizing the propositional complementation (represented at the surface level by various conjunctions linking the clauses) – groups of affirmative, imperative and interrogative verbs are distinguished there; further, (Kettnerová, 2009) deals with decomposition of the propositional complementation into so-called theme and dictum.

Alternation-Based Model of the Lexicon

Emphasizing syntactic criteria in the examination of valency results in the delimitation of specific valency frames for the use of the verbs with very similar (or the same) meanings but different syntactic structures, e.g., although the pairs such as *naložit vůz.*PAT *senem.*EFF *[to load the wagon.*PAT *with hay.*EFF*]* vs. *naložit seno.*PAT *na vůz.*DIR3 *[to load hay.*PAT *onto the wagon.* DIR3*]* či *vyběhnout kopec.*PAT *[to climb a hill]* vs. *vyběhnout na kopec.*DIR3 *[to climb up a hill]* differ in their syntactic structure (and thus are described by different valency frames), their semantic similarity is obvious and should be considered in the lexicon.

Let us mention so-called diathesis alternation in this context. Various kinds of diatheses (e.g., passive diathesis, deagentive diathesis, recipient diathesis, resultative diathesis or mediopassive diathesis) are characterized by different sentence structures; nevertheless, separating valency frames appears to be redundant with respect to regularity of changes in the syntactic structures affected (information about the potential of a verb to undergo a given diathesis alternation is sufficient here). We leave aside rich discussions concerning the question whether two sentences which differ from each other only in diathesis have the same meaning or not (e.g., sentences with the same lexical elements but differing in active / passive verb forms).

Similar method can be used also for the description of valency characteristics of

[19] At least some of them should be mentioned here, e.g., the above mentioned approach of B. Levin (1993), see also footnote 20, which was used in the VerbNet project, and semantic classification of verbs in the FrameNet project. In the Czech language, the classification of verbs in (Daneš – Hlavsa, 1987) is the most inspiring one.

verbs with different syntactic structure in case these structural (sometimes also semantic) shifts are systematic enough to be captured by linguistic rules. We use the term *alternations*[20] as a general term covering various phenomena that result in changes (of different types) in a valency structure of verbs.

The alternation-based model of the *VALLEX* lexicon and its logical structure was designed in (Žabokrtský, 2005). It enables systematic and efficient capturing of regular shifts of senses with separate verbs. This model of the lexicon is introduced in Section B.2 here.

The alternation-based model of the lexicon is characterized by two components – the data component and the grammatical component. The *data component* consists of lexemes that associate separate lexical forms and lexical units, roughly speaking, 'a given word in a given sense', see Chapter C. The *grammar component* describes the rules for various kinds of alternations (changes in valency structure); these changes may be manifested by:

- potential changes in a verb form;
- potential changes in a valency frame (i.e., changes in the number of valency complements and their functors, changes of the level of obligatoriness of complements and changes in their morphemic realization);
- possible shifts in meaning.

Section B.2 presents the initial classification of elementary types of alternations in *VALLEX* – so-called *syntactic alternations* (later referred to as *grammatical* alternations; especially various types of diatheses and reciprocalization, which are amply dealt with in Czech grammar books and theoretical articles, see e.g. Daneš et al., 1987; Grepl – Karlík, 1998; Panevová, 1999) and *semantic alternation* (e.g., cause co-occurrence, positive or negative, see Daneš, 1985) and the relations between them.

Classification of alternations has been further studied in relation to *VALLEX*; the more elaborate classification as well as the possibility of adequate capturing of alternations in a valency lexicon is presented in Section B.3.[21] Here alternations (or diatheses, in the terms used in Kettnerová – Lopatková, 2009b) are understood as changes in valency structure related to different mappings between valency slots and individual participants of a (shared) generalized situation (also type situation, see esp. Uspenskij, 1977). A basic typology of potential changes in valency structure is introduced there – *g-diathesis* and *s-diathesis* are distinguished. Roughly speaking, g-diatheses connect

[20] A significant contribution to the examination of various types of alternations was made by B. Levin (1993) who introduced the term 'alternation'. Based on an extensive list of alternations of various types she suggested a rich system of semantic classes. Although her work deals with English verbs, a lot of parallels with syntactical behavior of Czech verbs can be found there.

[21] In addition to already mentioned work, our inspiration comes from Czech and Slovak linguistics (see Section B.3), which was strongly influenced by Russian linguistics (esp. Apresjan, 1974; Cholodovič, 1970; Chrakovskij, 1977). The terms *hierarchization*, *diathesis* or *conversion* are used for similar concepts in these works.

pairs of related constructions characterized by changes in morphological verb forms (unmarked vs. marked form with respect to the category of voice) and changes typically limited to a choice of a particular valency member for the subject syntactic position - the mapping between semantic participants of a generalized situation and valency slots remains unchanged, see Figure 5. On the other hand, s-diatheses associate pairs of constructions that are characterized by changes in number and type of valency slots, while the (generalized) situation remains unchanged, see Figure 6; moreover, verbs are not morphologically marked with regard to voice.

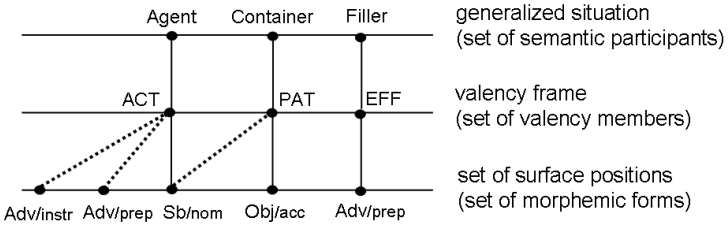


Figure 5: Mapping between semantic participants of a generalized situation and their surface syntactic positions for passive diathesis as a typical g-diathesis (for the verb *naložit [to load]*) (adopted from Kettnerová – Lopatková, 2009b).

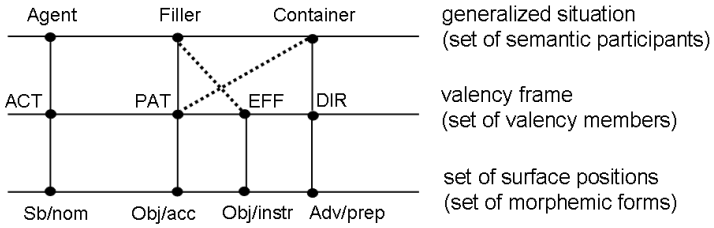


Figure 6: Mapping between semantic participants of a generalized situation and their surface syntactic positions for Container-Filler diathesis (for the verb *naložit [to load]*) (adopted from Kettnerová – Lopatková, 2009b).

We proposed a method of representation of these changes in *VALLEX*. In the case of g-diatheses, the changes in valency frames are regular enough to be treated in the form of general rules (in the grammar component) and as a single verbal lexical unit (for both syntactic constructions), marked with the possibility of a particular type of diathesis. On the other hand, for s-diatheses, separate lexical units are established and interlinked with general rules identifying a relevant type of s-diathesis as the changes in valency structure of verbs are diverse even within an individual type of s-diathesis (based on corpus evidence).

If separate lexical units are interlinked by alternation rules, it is possible to provide the valency information at different levels of compactness, e.g., according to the type of application for which the lexicon is intended.

It should be mentioned here that the conception of alternation-based model of the lexicon and its formal structure have been completed. The technology has been developed and implemented, too (Žabokrtský, 2005). So far, only a limited number of phenomena has been covered in the *VALLEX* data component (mainly those connected with reflexivity); see also Concluding Remarks in Section 6.

4 Current Concept of the Valency Lexicon: *VALLEX, Version 2*

The first version of the lexicon called *VALLEX 1.0* (and its quantitatively extended version *VALLEX 1.5*) treated the aspectual counterparts of verbs as separate entries (interlinked with a reference only). Such treatment of aspectual counterparts as separate units does not comply with the theoretical concept of FGD, which considers aspect to be a grammatical category and aspectual counterparts ('aspectual pairs' in common terminology) to be different realizations of one lexeme, see (Panevová et al., 1971).

This concept is reflected more adequately in the current version of the lexicon. *VALLEX, version 2*[22] treats valency characteristics of aspectual counterparts within one lexeme, which is represented by one lexicon entry only. The structure of *VALLEX, version 2* is described in detail in the introduction to the printed version of the lexicon, included in Chapter C here, and also on the Help page in HTML format on the lexicon website. At the topmost level, *VALLEX* is formed by lexemes – a *lexeme* is an abstract unit that associates a formal component (a set of all *lexical forms* of a given lexeme represented by their lemma(s)) and a semantic component represented by a set of individual *lexical units* (LU) – denoted as 'lexie' or elementary lexical units in Czech terminology, see Chapter C. The structure of a lexicon entry is shown in Figure 7.

VALLEX, version 2 describes the valency behavior of 2730 Czech lexemes, which comprise 6460 lexical units – roughly speaking, 'given verbs in given senses'. If the perfective and imperfective forms were counted separately, the number of verbs would grow up to 4250. The main criteria for the selection of verbs in *VALLEX* lexicon was their frequency in Czech National Corpus (ČNK) – during the first step, approximately 2500 verb lemmas with the highest rank were selected, after that each of the selected verbs was completed by its aspectual counterparts (if they are not already present in the list of the most frequent verbs), and occasionally also iterative counterparts.[23]

[22] In the following text the versions of the lexicon *VALLEX 2.0* and *VALLEX 2.5* are not strictly distinguished - both these versions have the same structure and describe the same number of verbs; version *2.5* underwent extensive and semi-automatic checking for correctness and consistency when the printed version of the lexicon was being prepared.

[23] Only aspectual counterparts formed by suffixes and rare aspectual suppletive pairs were treated this way. The reason for not including also aspectual counterparts formed by prefixes is purely practical; it is the unclear status of prefixed aspectual counterparts (esp. the selection of the proper counterpart if more prefixed counterparts exist). The possibility of linking prefixed perfective verbs to their non-prefixed imperfective counterparts is mentioned also in Concluding Remarks in Section 6.

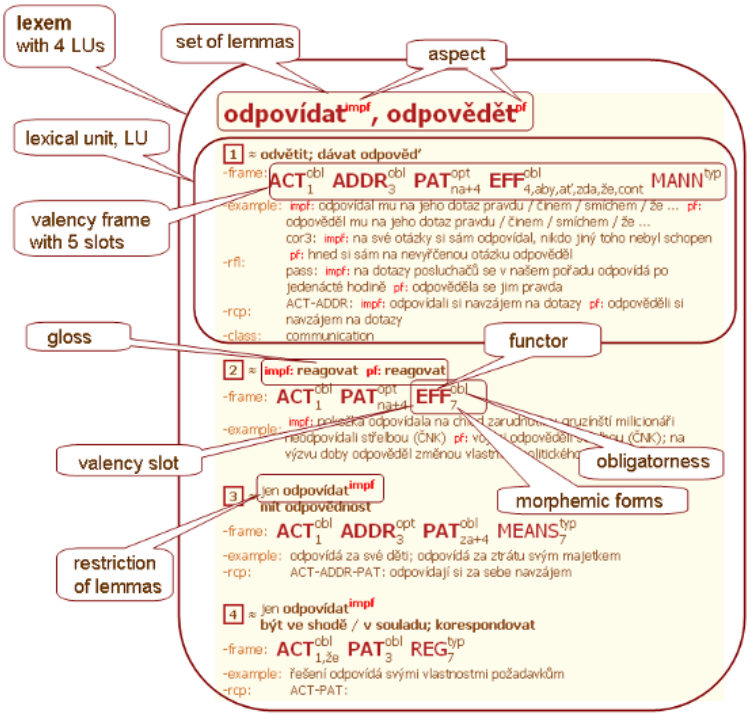


Figure 7: The structure of a lexicon entry in *VALLEX, version 2* (adopted from the Help page in HTML format).

The lexicon provides information on valency structure of Czech verbs in their particular senses, i.e., for particular lexical units. Each LU is specified with the use of glosses and examples. The core information on an individual LU is recorded in a form of a valency frame – a *valency frame* consists of a set of valency complementations, each of them being characterized (i) by its functor (type of syntactico-semantic relation between a verb and its complementation) and (ii) by its level of obligatoriness; in addition, (iii) possible morphemic forms are listed if these forms are determined by the verb's government. This obligatory information is accompanied with other syntactic or syntactico-semantic characteristics such as grammatical control, the type of reflexivity, possible reciprocal use or syntactico-semantic class of verbs.

The formal data structure of the lexicon and its technological and technical aspects are described in (Žabokrtský, 2005).

VALLEX 2.5 was published on the website of the Institute of Formal and Applied Linguistics at the Faculty of Mathematics and Physics, Charles University in Prague at the end of 2007.[24] Since the early 2008 a printed version has been available, too, published by Karolinum Press, the publishing house of Charles University in Prague.

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

5 Formal Modeling of Natural Language: Valency as Core Syntactic Information

The notion of valency is one of the basic concepts of dependency-based grammars. Valency characteristics of a verb represent core syntactic information determining sentence structure, see e.g. (Sgall, 1998, 2006); *valency (dependency) syntax* is sometimes used. Thus it is natural to use valency information also in formal modeling of syntactic structures of natural languages and their syntactic analysis.

Analysis by reduction in valency (dependency) syntax

The proposed method of *(dependency) analysis by reduction* is an elementary method for discovering syntactic structures of natural languages (and particularly languages with free word order).[25] Analysis by reduction (described in Section D.1. is based on a stepwise simplification of an analyzed sentence and makes it possible to define formal dependency relations between particular sentence members. While the basic operation in constituent-based approaches is the decomposition of the sentence into continuous parts representing simplified structures (phrases), in analysis by reduction it is possible to determine dependency relations between the sequences of words leaving aside the word order (at least to a certain degree; although at the same time the word order is not completely ignored). The principles of analysis by reduction can be summed up in the following observations:

1. The fact that a certain word (or sequence of words) can be deleted implies that this word or sequence of words *depends in analysis by reduction* on one of the words (or sequences of words) retained in the simplified sentence; the latter being called *governing word(s) in the reduction.*
2. Two words or sequences of words can be deleted in an arbitrary order if and only if they are mutually *independent in analysis by reduction.*
3. Certain sequences of words have to be deleted in a single step (taking into account the principles mentioned below). Even in such cases it is usual to determine governing and dependent words in dependency analysis. In such a case, it is necessary to define special rules for particular language phenomena

In the course of stepwise reduction of the analyzed sentence it is necessary to apply certain elementary principles:

- to preserve syntactic correctness of the sentence;
- to preserve lemmas (lexicon entries) and selected morphological tags (sets of morphological categories characterizing a given occurrence of the word);

[25] The fact that we are interested in a *formal model* of analysis must be stressed here, not in a psycholinguistic model which is to explain the process of understanding the sentences of natural language in the human mind.

- to preserve the senses of original words in the sentence (a sense is represented by a valency frame here);
- to preserve the completeness of the sentence at the layer of deep syntax (and especially to preserve the information about all inner participants and obligatory valency complementations of all non-reduced words in the sentence).

The method of analysis by reduction makes it possible to extract the dependency and especially valency relations in a sentence on the basis of potential order of the reductions of separate words or sequences of words. It is important especially for languages such as Czech, where the dependency structure cannot be extracted directly from the word order. Word order reflects topic-focus articulation and thus it carries deep-syntactic information (Hajičová et al., 1998). Changes in word order are not necessarily accompanied by changes in the dependency structure of the sentence; however, sentences differing from each other only in their word order cannot be considered as synonymous. So analysis by reduction enables us to study dependency relations and word order to a certain extent independently of each other.

The article in Section D.1 focuses on clarification of the relations between the analysis by reduction and the dependency-based representation of a sentence structure, see e.g. (Plátek – Holan, 2004). It shows that principles 1 and 2 of analysis by reduction can be used to model endocentric constructions, especially lexical words and their optional free modifications: words (or sequences of words) behaving like governing words in the analysis by reduction correspond to modified/governing words (or sequences of words) in the dependency analysis of a sentence while words (or their sequences) dependent in the reduction corresponds to a modifying/dependent words (or sequences of words) in the sentence.

The paper also analyzes so-called *reduction components* – sequences formed by words which have to be processed in one reduction step, see point 3. The reduction components correspond to exocentric constructions – they model for example formemes, i.e., word sequences forming separate sentence members (e.g., prepositional groups or analytical verb forms, see Sgall et al., 1986b; Sgall, 1998); determining a governing word in such cases is guided by the rules of a rather technical character (which may differ in various applications). However, another phenomenon is far more interesting: reduction components can adequately model a valency structure of verbs and other lexical words. Each frame-evoking unit must be deleted together with all its valency complementations in a single reduction step; none of these complementations can be deleted earlier (otherwise the principle of completeness mentioned above is violated). We adopt the principle of analogy at the level of parts of speech – this principle, which was proposed in (Sgall et al., 1986b) is used for the specification of the direction of dependency relations.

Restarting automata as a formal device for modeling analysis by reduction

The dependency-based Functional Generative Description was originally modeled as a generative system. A serial composition of pushdown and finite automata-transducers was proposed as a model of the translation component of FGD; see especially (Sgall, 1967; Sgall et al., 1969; Plátek – Sgall, 1978). A model of the generative component of FGD (i.e., the component, which constitutes the deep syntactic (tectogrammatical) representation of a sentence) was later described in detail in (Petkevič, 1995). This model was constructed as a push-down automaton, too.

In 1980s a system of translation schemes was designed, which made the interpretation in both directions possible; i.e., it worked as both a generative and an analytical system (Plátek, 1982). Similar description of the FGD model can be found in (Sgall et al., 1986a), where it is modeled as a set of several context-free languages corresponding to separate layers of language description.

Analysis by reduction provided a crucial motivation for a new formal model of FGD based on the novel concept of *restarting automata*. Only an elementary and informal description of restarting automata is provided here – their formal description and detailed typology can be found in the ample bibliography devoted to these automata, their properties, and hierarchies of the languages accepted by them, see e.g. (Jančar et al., 1999; Otto, 2006), and the works cited there. Analysis by reduction can be adequately modeled by restarting automata which work with *input language* (language of an input sentence) and *characteristic language* (input language enriched with grammatical categories describing the sentence structure); see especially (Messerschmidt et al., 2006; Mráz et al., 2007; Plátek – Otto, 2008).

In the presented work a particular model of a restarting automaton is formally defined in Section D.2. This type of restarting automaton, modeling analysis by reduction (henceforth referred to as RA; $4 - LRL$-automaton in D.2), is a formal device – a non-deterministic machine with a finite-state control unit, a finite characteristic vocabulary and a head which can read and process the symbols (words) of the sentence on a flexible working tape, marked by special symbols (the end-markers), see Figure 8.

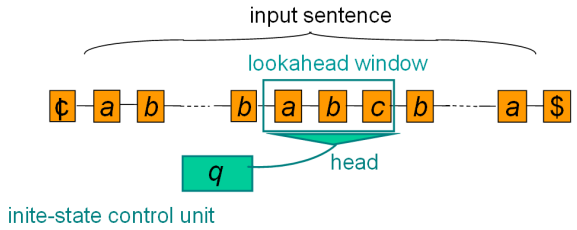


Figure 8: A diagram of RA restarting automaton.

This type of automaton starts the computation over an input sentence in the initial state with its head placed on the left end of the tape. During the computation, RA performs the following operations according to its transition relation:

MVR/MVL: move right / left operations change the state of RA and shift the window;

Rewrite(v): a rewrite step shortens the sentence on the working tape, i.e., rewrites the sequence of words in the lookahead window by a shorter sequence v;

Restart: the head moves to the left end of the tape and reenters the initial state;

Accept/Reject: the automaton accepts/rejects the input sentence on the tape.

A computation of a restarting automaton consists of cycles (see Figure 9); the input sentence is processed – according to the transition relation of the automaton, the head reads the words on the tape, moves right or left and rewrites / shortens the sentence on the tape – until the sentence is accepted / rejected or until a restart operation is performed. Then the position of the head as well as the inner state of the control unit is 'forgotten' and the head starts processing the (already shortened) sentence from the beginning in a new cycle.

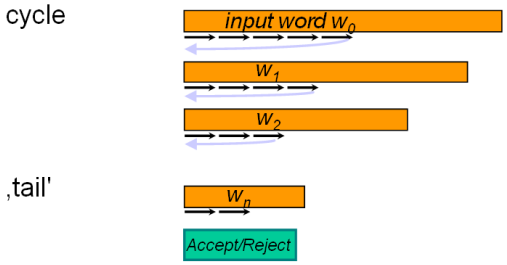


Figure 9: The computation of a restarting automaton consisting of cycles

The essential property of this type of restarting automaton is an *error preserving property* – for any accepting computation, there is a sentence from the characteristic language on the tape before and after each cycle. For modeling the analysis by reduction, a stronger property, a *correctness preserving property*, is usually demanded – this requirement guarantees that each computation of a restarting automaton over a sentence in the characteristic language is an accepting computation.

Modeling analysis by reduction (and consequently also syntactic analysis) with the use of restarting automata reflects the paradigm of FGD better than earlier models based on push-down automata:

- Restarting automaton RA models adequately the syntactic relations determined by valency characteristics of lexical words. It makes it possible to perform several rewrite steps in a single cycle. This feature is used for processing reduction components (consisting of frame-evoking words and their valency complementations) – the processing of a single verb (or noun, adjective or adverb) and its valency complements is modeled in one computational cycle. Therefore RA reflects the *complete valency structures* as it is understood in the concept of valency

syntax; that distinguishes RA significantly from the models based on pushdown automata, which model *syntactic pairs* consisting of a governing and a dependent word.

- Restarting automaton RA makes it possible to capture the concept of *lexikalization* – the approach characteristic for dependency-based language description, which collects essential linguistic information in a lexicon (reference should be made here to at least categorial grammars, see Ajdukiewicz, 1935, and the concept of *lexicalized tree adjoining grammar*, see e.g. Abeillé – Rambow, 2000, based on the *tree adjoining grammars*, Joshi, 1985; within FGD compare for example Sgall, 1998).
- Restarting automaton RA reflects non-local behavior of languages with free word order – rewrite steps in such general models of automata are not restricted to the continuous substring of an input sentence (Plátek et al., 2005; Plátek – Otto, 2008); they can reduce several symbols with distant word-order positions (stored as discontinuous strings on a working tape). Thus RA can process words (and their complementations) with unbounded positions in a sentence as well as words forming non-projective (surface) constructions.
- A restarting automaton working in cycles models recursive properties of a language appropriately – first, the deepest embedded language constructions are processed, which results in the simplification of an analyzed sentence; then the language constructions embedded in such simplified sentence are processed; after each simplifying operation a new cycle starts (i.e., the automaton restarts). The computation proceeds until the so-called core predicative structure is reached and accepted without any further restart (in the 'tail' of the computation, see Figure 9) or until the simplified sentence is rejected as an ill-formed sentence.

Functional Generative Description as a formal translation

A formal system for a description of a natural language is required to be able to describe the set of correct sentences of a language, the set of potential deep syntactic (tectogrammatical) representations of the sentences in the given language and the relations between these two sets reflecting the relations of representation (and thus capturing synonymy and ambiguity in the language, see Plátek, 1982; Sgall et al., 1986a).

In (Lopatková et al., 2008), Section D.2, here, a 4-level reduction system is defined. This reduction system represents a new formal frame for the modeling of FGD based on the principles of analysis by reduction.

A restarting automaton RA modeling FGD, hereinafter M_{FGD}, processes sentences over a characteristic vocabulary, which consists of all word forms of the natural language concerned as well as of all grammatical categories describing the sentence structures. The stratification based approach of FGD is manifested by the division of the characteristic vocabulary Σ into four (sub)vocabularies for separate language layers:

Σ_w ... vocabulary consisting of all correct word forms in the language; i.e., input

language of M_{FGD} automaton representing a layer of words of a given natural language;

Σ_m ... vocabulary for morphological analysis describing lemmas and their morphological tags, i.e., a morphological layer;

Σ_a ... vocabulary representing surface syntax, i.e., analytical layer;[26]

Σ_t ... vocabulary consisting of units describing deep syntactic characteristics of lexical words (especially lexical and valency information, functors, gramatemes and topic focus articulation), i.e., information from the tectogrammatical layer of a language description.[27]

The characteristic language of the M_{FGD} automaton is specified in Section D.2. This language captures a modeled sentence and its analysis at separate layers (with the help of $\Sigma_w, \Sigma_m, \Sigma_a$ a Σ_t vocabularies). A computation of M_{FGD} automaton is illustrated on particular Czech sentences there. Attention is paid to two essential linguistic phenomena. First, the processing of valency and free modifications, the principle of preserving the completeness of the input sentence being a crucial requirement here; the essential principles of analysis by reduction stated above (preservation of syntactic correctness, preservation of lemmas and tags, preservation of word senses and preservation of completeness) guarantee the adequate description of the complete deep syntactic (tectogrammatical) representation of a sentence. The other phenomenon elaborated here in detail is the representation of both surface and deep word order (including non-projective constructions, which may occur in the surface representation of a sentence, see especially Holan et al., 2000; Zeman, 2004; Hajičová, 2006).

The M_{FGD} restarting automaton accepts exactly all correct (well-formed) sentences of the modeled natural language together with their (disambiguated) representations at all layers; it rejects the sentences which do not belong to this natural language or which are characterized by incorrect representation in any of the descriptive layers.

The formal relation between the projection of the processed sentence into the layer represented byΣ_w vocabulary (i.e., in the input language of M_{FGD}) and the projection into the layer represented by Σ_t vocabulary defines the *characteristic relation*. This relation models the relation of representation, i.e., the relation between the set of sentences in a given natural language and the set of deep syntactic (tectogrammatical) representations of the sentences. The characteristic relation is interpreted as a formal translation from the language of correct (well-formed) sentences into the language of tectogrammatical representation, (i.e., formal analysis) or as a translation from the tectogrammatical language into the language of correct sentences (i.e., formal synthesis).

A formal model of FGD is further studied in (Plátek – Lopatková, 2007), where this system is classified among formal translation systems while the emphasis is put on the connection between the formal models and their linguistic content. Another con-

[26] The question of theoretical adequacy of this layer of FGD is not taken into consideration here.

[27] Subvocabulary Σ_w is the vocabulary of the input language of M_{FGD}, vocabulary $\Sigma = \Sigma_w \cup \Sigma_m \cup \Sigma_a \cup \Sigma_t$ is the vocabulary of the characteristic language of M_{FGD} automaton.

tribution to adequate model of FGD is represented by a model of restarting automata with structured output (Plátek et al., 2010). Here a restarting automaton is treated as a transducer which processes input strings from characteristic language and yields tree structures, from which dependency trees can be derived (namely DR-trees, see e.g. Holan et al., 1998.

The framework of analysis by reduction (and its modeling by restarting automata) makes it possible to define detailed rules for particular linguistic phenomena. When processing valency-based relations and elementary word order phenomena (including surface non-projectivity) it is possible to work with a restricted type of restarting automaton in which rewriting operations are reduced to deletions (i.e., certain symbols from the working tape are simply deleted during the rewrite steps). Other constructions, for example numeral constructions, see Section D.1, and especially coordination constructions, require a more general model of a restarting automaton with 'real' rewriting.

It should be noted that coordination and appositional constructions have not been considered in formal description based on restarting automata so far because they work with units of constituent character (in the sense of constituent-based approaches, see (Plátek et al., 1985; Sgall, 1998). They go significantly beyond the straightforward concept of purely dependency-based approaches. Nevertheless, at present even these constructions are being treated gradually within the framework of analysis by reduction and the paradigm established by restarting automata.

6 Concluding Remarks

Practical use of *VALLEX* in building other lexicons and in NLP applications

The *VALLEX* lexicon has been designed with strong emphasis on the exactness and linguistic adequacy as well as consistency of valency description for a large number of verbs. Lexicon entries were processed manually, stress was laid on corpus evidence and dictionary material. The manual phase was followed by extensive automatic, semi-automatic and manual checking. From the very beginning, *VALLEX* was intended for both people as language users and for computational processing the Czech language in applied tasks such as machine translation, text searching, etc.

- Valency lexicon *VALLEX* is used by more than 150 registered users, mostly at Czech universities but also at several international universities and research centers. Most of the licenses were issued to various workplaces at the Faculty of Mathematics and Physics and the Faculty of Arts at Charles University in Prague, the Faculty of Information Technologies at Masaryk University and the Institute of the Czech Language at the Academy of Sciences of the Czech Republic; dozens of licenses for international institutions involve, e.g., the Ohio State University, Saarland University (Universität des Saarlandes), University in Zagreb

(Sveučilište u Zagrebu) or INALCO (Institut National des Langues et Civilisations Orientales) and LaLIC (Language, Logiques, Informatique, Cognition) at Paris-Sorbonne University (l'Université Paris-Sorbonne).

At least some of the applications making use of the lexicon data or its technological processing methods should be mentioned here, too.

- The logical structure of *VALLEX 1.0* and its technological implementation (especially the data format, XML representation, conversion and validation scripts of Z. Žabokrtský) are intensively used for building the *VerbaLex* lexicon, see (Hlaváčková – Horák, 2005, 2006; Hlaváčková, 2008).
- The structure of the *VALLEX* lexicon and the experience gained during the processing of verbs were utilized in the course of the compilation of the valency lexicon of English verbs *EngVallex*, see (Cinková, 2006), which was built on the basis of the *PropBank Lexicon* (Kipper et al., 2004; Palmer et al., 2005). The *EngVallex* lexicon is used for manual annotation of the tectogrammatical representation of English in the Prague Czech-English Dependency Treebank (work in progress) and is manually interlinked with the *PDT-VALLEX* lexicon (Šindlerová – Bojar, 2009).
- Some principles and experience gained from building the *VALLEX* lexicon were also used for developing a Swedish-Czech valency dictionary, see (Cinková – Žabokrtský, 2005a,b; Cinková, 2009).
- Valency theory has been developed primarily for verbs. Valency characteristics of deverbal nouns and adjectives reflect to some degree valency properties of base verbs. Deverbal nouns inherit valency frames of the base verbs to a certain extent; see especially (Panevová, 2000, 2003). What is significant here is the type of derivation, i.e., whether the derivation is syntactic or lexical.
  A pilot study concerning the possibilities of predicting a valency frame of a noun on the basis of the type of derivation (especially the type of a suffix) and the valency frame of the base verb in *VALLEX 1.0* was summarized in (Lopatková et al., 2002a), later it was elaborated thoroughly in the articles and PhD. thesis of V. Kolářová-Řezníčková, see especially (Kolářová, 2005, 2006).
- A random selection of 109 verbs from *VALLEX 1.0* was used for a lexical sampling experiment resulting in the *VALEVAL* corpus, see (Bojar et al., 2005), here also Section A.3. For each of these verbs, 100 sentences were extracted from ČNK and appropriate LUs from the lexicon were manually assigned to them. So-called golden data – *golden VALEVAL* – represent a set of sentences in which the annotators agreed on the assigned LU (and thus sense). The pairwise inter annotator agreement was about 75%, which is comparable to the results achieved for significant lexicons of English verbs, e.g., *PropBank Lexicon* (based on an oral statement of M. Palmer).
  The corpus consisting of sentences with explicitly disambiguated words is a fundamental prerequisite for the development of a tool for 'word sense disambigua-

tion', i.e., a tool which automatically assigns the senses to particular occurrences of words in a text.

- The *VALEVAL* corpus was used for training the tools for word sense disambiguation based on various machine learning methods, see especially (Semecký – Podveský, 2006; Semecký, 2007), also Section A.3 here. It was proved that the treatment of verbs in the *VALLEX* lexicon is so consistent that it is possible to train a tool which is able to recognize the correct valency frame for a given verb occurrence with the success rate of 77.2% (comparable to the baseline 60.7% or the assigning the most frequent valency frame to each occurrence of the verb), see the works by J. Semecký cited above.

Further development of the *VALLEX* lexicon

Extensive data processing resulting in the *VALLEX* lexicon shows valency as both a syntactic (combinatorial) and a lexicographical phenomenon. It shows that a dictionary-based approach to capturing valency entails new theoretical problems, which require further and more detailed linguistic examination.

The most difficult task from a lexicographical point of view is still the delimitation of separate senses of verbs. In the *VALLEX* lexicon, emphasis is put on syntactic criteria because they are more explicit (and testable) than the criteria used for deeper layers of description. However, at the same time it is undoubtedly necessary to take account of semantics, too.

Questions of sense specification are closely related to the very interesting and highly relevant concept of alternations – the ability of verbs to determine various syntactic structures while the basic meaning is preserved. So far, basic types of phenomena that should be captured in the lexicon have been classified. The current logical structure of the data already reflects the requirements for the alternation-based model.[28] At the present stage it is necessary to build the grammar component of the lexicon that captures detailed description of the rules for particular types of alternations (on the basis of existing studies, see Section 3) and it is also necessary to indicate systematically possible alternations for particular lexical units in the data component of the lexicon.

The general concept of the alternations also enables grouping together of the verbs and their prefixed derivates and, in particular, linking imperfective verbs to their prefixed perfective counterparts at least at this level.

Another interesting question is the possibility of enhancing *VALLEX* with deep semantic information - semantic classes and semantic roles, especially the information stemming from the existing data resources. There are pilot studies focusing on several groups of verbs, namely verbs of communication, mental action, exchange, motion, transport and psych verbs Kettnerová et al., 2008; Kettnerová – Lopatková, 2009a). These studies examine the possibilities of potential interlinking these syntactically and semantically heterogeneous groups of verbs with the elaborated and highly appreciated

[28] At the technological level, the alternation-based model of the lexicon has been implemented by Z. Žabokrtský; see Žabokrtský, 2005.

network of semantic frames *FrameNet*,[29] in which English verbs, nouns, adjectives and adverbs are treated (Ruppenhofer et al., 2006). Such interlink could also enable the classification of Czech verbs into more subtle syntactically and semantically coherent classes using semantic frames. Let us mention also another challenging possibility to enhance the existing lexicon (and to verify its analysis of verb lexemes), namely the method of *corpus pattern analysis* proposed in (Hanks, 2004, 2010),[30] which consists in the analysis of prototypical syntagmatic patterns based on verb occurrences in a large corpus, together with the semantic types of the complementations.

[29] See also Section 3, footnote 16.
[30] See also Section 3, footnote 17.

Bibliography

Abeillé, A. – Rambow, O. (eds.) (2000): *Tree Adjoining Grammars: Formalisms, Linguistic Analysis and Processing.* Stanford, Center for the Study of Language and Information.

Ajdukiewicz, K. (1935): Die syntaktische Konnexität. *Studia Philosophica.* I, p. 1–27.

Apresjan, J. D. (1974): *Leksicheskaja semantika. Sinonimicheskie sredstva jazyka.* Moskva, Nauka.

Bejček, E. (2009): Automatické přiřazování valenčních rámců a jejich slévání. In Vojtáš, P. (ed.) *Proceedings of ITAT 2009*, p. 9–14, Seňa, Slovakia. PONT s.r.o.

Bojar, O. – Semecký, J. – Benešová, V. (2005): Testing VALLEX Consistency and Experimenting with Word-Frame Disambiguation. *The Prague Bulletin of Mathematical Linguistics.* 83, p. 5–17.

Cinková, S. (2006): From PropBank to EngValLex. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, p. 2170–2175, Paris. ELRA.

Cinková, S. (2009): *Words that Matter: Towards a Swedish-Czech Colligational Dictionary of Basic Verbs.* 2 / *Studies in Computational and Theoretical Linguistics.* Malostranské nám. 25, 118 00 Praha 1, UFAL.

Cinková, S. – Žabokrtský, Z. (2005a): Swedish-Czech Combinatorial Valency Lexicon of Predicate Nouns: Describing Event Structure in Support Verb Constructions. In Kiefer, F. – Kiss, G. – Pajzs, J. (eds.) *Proceedings of the 8th International Conference on Computational Lexicography COMPLEX*, p. 50–59, Budapest.

Cinková, S. – Žabokrtský, Z. (2005b): Treating Support Verb Constructions in a Lexicon: Swedish-Czech Combinatorial Valency Lexicon of Predicate Nouns. In Erk, K. – Melinger, A. – Schulte im Walde, S. (eds.) *Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, p. 22–27, Saarbrücken.

Daneš, F. (1971): Větné členy obligatorní, potenciální a fakultativní. *Miscellanea Linguistica.* p. 131–138.

Daneš, F. (1985): *Věta a text.* Praha, Academia.

Daneš, F. – Grepl, M. – Hlavsa, Z. (eds.) (1987): *Mluvnice češtiny 3.* Praha, Academia.

Daneš, F. – Hlavsa, Z. (1987): *Větné vzorce v češtině.* Praha, Academia. (co-authors: Jirsová, A. – Macháčková, E. – Prouzová, H. – Svozilová, N.).

Fellbaum, C. (ed.) (1998): *WordNet: An Electronic Lexical Database.* Cambridge MA, MIT Press.

Fillmore, C. J. (1968): The Case for Case. In Bach, E. – Harms, R. T. (eds.) *Universals in Linguistic Theory*, p. 1–88. New York: Holt, Rinehart and Winston.

Fillmore, C. J. (1969): Types of Lexical Information. In Kiefer, F. (ed.) *Studies in syntax and semantics*, p. 109–137. New York: Kluwer Academic Publishers.

Fillmore, C. J. – Johnson, C. – Petruck, M. R. L. (2003): Background to FrameNet.

International Journal of Lexicography. 16, 3, p. 235–250.

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

Hajič, J. (1998): Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In Hajičová, E. (ed.) *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová*, p. 106–132. Prague: Karolinum Press.

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

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*, Vol. 9, p. 57–68. (co-authors: Panevová, J. – Urešová, Z. – Bémová, A. – Kolářová, V. – Pajas, P.).

Hajičová, E. (2006): K některým otázkám závislostní gramatiky. *Slovo a slovesnost*. 67, 1, p. 3–26.

Hajičová, E. – Partee, B. H. – Sgall, P. (1998): *Topic-Focus Articulation, Tripartite Structures, and Semantic Content*. Dordrecht, Kluwer.

Halliday, M. A. K. (1963): Some Notes on 'Deep' Grammar. *Journal of Linguistics*. 2, 1, p. 57–67.

Hanks, P. (2004): Corpus Pattern Analysis. In Williams, G. – Vessier, S. (eds.) *Euralex Proceedings*, I, p. 87–98, Lorient, France.

Hanks, P. (2010): *Lexical Analysis: Norms and Exploitations*. , MIT Press. (forthcoming).

Hanks, P. – Pustejovsky, J. (2005): A Pattern Dictionary for Natural Language Processing. *Revue Française de Langue Appliquée*. 10, 2, p. 63–82.

Havránek, B. (ed.) (1964): *Slovník spisovného jazyka českého*. Praha, Academia.

Hlaváčková, D. (2008): *Databáze slovesných valenčních rámců VerbaLex*. PhD thesis, Masarykova Universita, Brno.

Hlaváčková, D. – Horák, A. (2005): Transformation of WordNet Czech Valency Frames into Augmented VALLEX-1.0 Format. In *Human Language Technologies as a Challenge for Computer Science and Linguistics*, p. 310–313, Poznan. Wydawnictwo Poznańskie Sp. z o.o. with cooperation of Fundacja Uniwersytetu im. A. Mickiewicza.

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

Holan, T. et al. (1998): Two Useful Measures of Word Order Complexity. In Polguére, A. – Kahane, S. (eds.) *Proceedings of the Workshop 'Processing of Dependency-Based Grammars', COLING-ACL'98*, p. 21–28, Montréal, Quebeck. (co-authors: Kuboň, V. – Oliva, K. – Plátek, M.).

Holan, T. et al. (2000): On Complexity of Word Order. *Les grammaires de dépendance – Traitement automatique des langues (TAL)*. 41, 1, p. 273–300. (co-authors: Kuboň, V. – Oliva, K. – Plátek, M.).

Horák, A. (1998): Verb Valency and Semantic Classification of Verbs. In Sojka, P. et al. (eds.) *Proceedings of Text, Speech and Dialog International Conference, TSD'98*, p. 61–66, Brno.

Cholodovič, A. A. (1970): Zalog. Kategoria zaloga. In *Materialy konferencii*, p. 2–26,

Leningrad.

Chrakovskij, V. S. (ed.) (1977): *Problemy lingvisticheskoj tipologii i struktury jazyka.* Leningrad, Nauka.

Jančar, P. et al. (1999): On Monotonic Automata with a Restart Operation. *Journal of Automata, Languages and Combinatorics.* 4, 4, p. 287–311. (co-authors: Mráz, F. – Plátek, M. – Vogel, J.).

Joshi, A. (1985): Tree Adjoining Grammars: How Much Context-Sensitivity is Required to Provide Reasonable Structural Descriptions? In Dowty, D. (ed.) *Natural Language Processing*, p. 206–250. Cambridge: Cambridge University Press.

Karlík, P. (2000): Hypotéza modifikované valenční teorie. *Slovo a slovesnost.* 61, p. 170–189.

Kettnerová, V. (2008): Czech Verbs of Communication with respect to the Types of Dependent Content Clauses. *The Prague Bulletin of Mathematical Linguistics.* (to appear).

Kettnerová, V. (2009): Konstrukce s rozpadem tématu a dikta v češtině. *Slovo a slovesnost.* 70, 3, p. 163–174.

Kettnerová, V. – Lopatková, M. (2009a): Mapping Semantic Information from FrameNet onto VALLEX. In *FrameNet Masterclass and Workshop*, Milan. (contributed talk).

Kettnerová, V. – Lopatková, M. (2009b): 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*, p. 198–210, Bratislava, Slovakia. Slovenská akadémia vied.

Kettnerová, V. – Lopatková, M. – Hrstková, K. (2008): Semantic Classes in Czech Valency Lexicon: Verbs of Communication and Verbs of Exchange. In Sojka, P. et al. (eds.) *Proceedings of Text, Speech and Dialog International Conference, TSD 2008*, LNAI, p. 109–116, Berlin Heidelberg. Springer-Verlag.

Kipper, K. – Snyder, B. – Palmer, M. (2004): Extending a Verb-lexicon Using a Semantically Annotated Corpus. In *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC 2004, Workshop on Building Lexical Resources from Semantically Annotated Corpora*, Paris. ELRA.

Kolářová, V. (2005): Valence deverbálních substantiv: Některé specifické posuny v povrchových realizacích participantů. In Karlík, P. (ed.) *Sborník konference Korpus jako zdroj dat o češtině*, p. 113–125, Brno.

Kolářová, V. (2006): *Valence deverbativních substantiv v češtině.* PhD thesis, Univerzita Karlova v Praze, Praha.

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

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

Lopatková, M. (2001): *Homonymie předložkových skupin a možnost jejich automatického zpracování.* PhD thesis, Univerzita Karlova v Praze.

Lopatková, M. (2003a): Valency in the Prague Dependency Treebank: Building the Valency Lexicon. *The Prague Bulletin of Mathematical Linguistics.* 79–80, p. 37–60.

Lopatková, M. (2003b): *O homonymii předložkových skupin v češtině (Co umí počítač?).* Praha, Nakladatelství Karolinum.

Lopatková, M. – Panevová, J. (2004): Valence vybraných skupin sloves (k některým

slovesům dandi a recipiendi). In Hladká, Z. – Karlík, P. (eds.) *Čeština – univerzália a specifika, Sborník konference ve Šlapanicích u Brna*, Vol. 5, p. 348–356. Praha: Nakladatelství Lidové noviny.

Lopatková, M. – Panevová, J. (2006): Recent Developments in the Theory of Valency in the Light of the Prague Dependency Treebank. In Šimková, M. (ed.) *Insight into Slovak and Czech Corpus Linguistics*, p. 83–92. Bratislava: Veda.

Lopatková, M. – Panevová, J. (2007): Valence vybraných sloves pohybu v češtině (antonyma, nebo synonyma?). In Piper, P. (ed.) *Sbornik Matice srpske za slavistiku*, 71-72/2007, p. 105–115, Novi Sad. Matica srpska.

Lopatková, M. – Plátek, M. – Kuboň, V. (2005b): Modeling Syntax of Free Word-Order Languages: Dependency Analysis by Reduction. In Matoušek, V. – Mautner, P. – Pavelka, T. (ed.) *Proceedings of Text, Speech and Dialogue International Conference, TSD 2005*, 3658 / *LNAI*, p. 140–147, Berlin Heidelberg. Springer-Verlag.

Lopatková, M. – Plátek, M. – Sgall, P. (2008): Functional Generative Description, Restarting Automata and Analysis by Reduction. In Marušič, F. – Žaucer, R. (eds.) *Studies in Formal Slavic Linguistics. Contributions from Formal Description of Slavic Languages 6.5.*, Vol. 19 / *Linguistik International*, p. 173–190. Frankfurt am Main: Peter Lang Publishing Group.

Lopatková, M. – Řezníčková, V. – Žabokrtský, Z. (2002a): Valency Lexicon for Czech: from Verbs to Nouns. In Sojka, P. – Kopeček, I. – Pala, K. (eds.) *Proceedings of Text, Speech and Dialogue International Conference, TSD 2002*, 2448 / *LNAI*, p. 147–150, Berlin Heidelberg. Springer-Verlag.

Lopatková, M. – Žabokrtský, Z. (2003): Testování konzistence a úplnosti valenčního slovníku českých sloves. In Vojtáš, P. (ed.) *Proceedings of ITAT 2003*, p. 73–82, Košice. University of P. J. Šafárik.

Lopatková, M. – Žabokrtský, Z. – Skwarska, K. (2006): Valency Lexicon of Czech Verbs: Alternation-Based Model. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, Vol. 3, p. 1728–1733, Paris. ELRA.

Lopatková, M. et al. (2002b): Tektogramaticky anotovaný valenční slovník českých sloves. Technical Report TR-2002-15, ÚFAL/CKL MFF UK, Praha. (co-authors: Žabokrtský, Z. – Skwarska, K. – Benešová, V.).

Lopatková, M. et al. (2003): VALLEX 1.0 Valency Lexicon of Czech Verbs. Technical Report TR-2003-18, ÚFAL/CKL MFF UK, Prague. (co-authors: Žabokrtský, Z. – Skwarska, K. – Benešová, V.).

Lopatková, M. et al. (2005a): Valency Lexicon of Czech Verbs VALLEX: Recent Experiments with Frame Disambiguation. In Matoušek, V. – Mautner, P. – Pavelka, T. (eds.) *Proceedings of Text, Speech and Dialogue International Conference, TSD 2005*, 3658 / *LNAI*, p. 99–106, Berlin Heidelberg. Springer-Verlag. (co-authors: Bojar, O. – Semecký, J. – Benešová, V. – Žabokrtský. Z.).

Marcus, M. P. – Santorini, B. – Marcinkiewicz, M. A. (1993): Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*. 19, 2, p. 313–330.

Matthews, P. H. (1997): *The Concise Oxford Dictionary of Linguistics*. Oxford, Oxford University Press. (dictionary definition: Valency).

Messerschmidt, H. et al. (2006): Correctness Preservation and Complexity of Simple RL-Automata. In *Implementation and Application of Automata*, 4094 / *LNCS*, p. 162–172, Berlin

Heidelberg. Springer-Verlag. (co-authors: Mráz, F. – Otto, F. – Plátek, M.).

Mráz, F. – Plátek, M. – Otto, F. (2007): Free Word-Order and Restarting Automata. In *Pre-proceedings of LATA 2007*, p. 425–436, Taragona. Universitat Rovira I Virgili.

Otto, F. (2006): Restarting Automata. In Ésik, Z. – Martin-Vide, C. – Mitrana, V. (eds.) *Recent Advances in Formal Languages and Applications, Studies in Computational Intelligence*, Vol. 25, p. 269–303, Berlin. Springer-Verlag.

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

Pala, K. – Ševeček, P. (1997): Valence českých sloves. In *Sborník prací FFBU*, p. 41–54, Brno.

Palmer, M. – Gildea, D. – Kingsbury, P. (2005): The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*. 31, 1, p. 71–106.

Panevová, J. (1974-5): On Verbal Frames in Functional Generative Description I-II. *The Prague Bulletin of Mathematical Linguistics*. 22-23, p. 3–40,17–52.

Panevová, J. (1980): *Formy a funkce ve stavbě české věty*. Praha, Academia.

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

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

Panevová, J. (2000): Poznámky k valenci podstatných jmen. In Karlík, V. – Hladká, Z. (eds.) *Čeština – univerzália a specifika. Sborník konference ve Šlapanicích u Brna 17.-18. 11. 1998*, p. 173–180. Brno: Masarykova univerzita.

Panevová, J. (2003): K valenci substantiv (s ohledem na jejich derivaci). In Piper, P. (ed.) *Sbornik Matice srpske za slavistiku*, p. 29–36. Novi Sad: Matica srpska.

Panevová, J. – Benešová, E. – Sgall, P. (1971): *Čas a modalita v češtině*. 34 / *Philologica Monographi*. Praha, Acta Universitatis Carolina.

Panevová, J. – Skoumalová, H. (1992): Surface And Deep Cases. In *Proceedings of COLING'92*, p. 885–889.

Petkevič, V. (1995): A New Formal Specification of Underlying Structure. *Theoretical Linguistics*. 21, 1, p. 7–61.

Plátek, M. (1982): Composition of Translation with D-trees. In Horecký, J. (ed.) *Proceedings of the 9th International Conference on Computational Linguistics, COLING'82*, p. 313–318, Prague. Academia.

Plátek, M. – Holan, T. (2004): Závislostní a složkové modelování syntaxe jazyků. In Obdržálek, D. – Stanclová, J. (eds.) *Malý informatický seminář, MIS 2004*, p. 115–150, Praha. Matfyzpress.

Plátek, M. – Lopatková, M. (2007): Funkční generativní popis a formální teorie překladů. In Vojtáš, P. (ed.) *Proceedings of ITAT 2007*, p. 3–14, Košice. University of P. J. Šafárik.

Plátek, M. – Mráz, F. – Lopatková, M. (2010): Restarting Automata with Structured Output and Functional Generative Description. In *Proceedings of LATA 2010*, LNCS, Berlin Heidelberg. Springer-Verlag.

Plátek, M. – Otto, F. (2008): A Two-Dimensional Taxonomy of Proper Languages of

Lexicalized FRR-Automata. In *Pre-proceedings of LATA 2008*, Taragona. Universitat Rovira I Virgili.

Plátek, M. – Sgall, P. (1978): A Scale of Context-Sensitive Languages: Applications to Natural Language. *Information and Control*. 38, 1, p. 1–20.

Plátek, M. – Sgall, J. – Sgall, P. (1985): A Dependency Base for a Linguistic Description. In Sgall, P. (ed.) *Contribution to Functional Syntax, Semantics and Language Comprehension*, 16 / *Linguistic and Literary Studies in Eastern Europe*, p. 63–97. Prague: Academia.

Plátek, M. et al. (2005): O roztržitosti a volnosti slovosledu pomocí restartovacích automatů. In Vojtáš, P. (ed.) *Proceedings of ITAT 2005*, p. 145–156, Košice. University of P. J. Šafárik. (co-authors: Mráz, F. – Otto, F. – Lopatková, M.).

Ruppenhofer, J. et al. (2006): *FrameNet II: Extended Theory and Practice*. Berkeley, University of California. http://framenet.icsi.berkeley.edu/book/book.html, (co-authors: Ellsworth, M. – Petruck, M. R. L. – Johnson, C. R. – Scheffczyk, J.).

Sarkar, A. – Zeman, D. (2000): Automatic Extraction of Subcategorization Frames for Czech. In *Proceedings of the 18th International Conference on Computational Linguistics, COLING 2000*, p. 691–697, Saarbrücken, Germany.

Semecký, J. (2007): *Verb Valency Frames Disambiguation*. PhD thesis, Charles University in Prague, Prague.

Semecký, J. – Podveský, P. (2006): Extensive Study on Automatic Verb Sense Disambiguation in Czech. In Sojka, P. – Kopecek, I. – Pala, K. (eds.) *Proceedings of Text, Speech and Dialog International Conference, TSD 2006*, 4188 / *LNAI*, p. 237–244, Berlin Heidelberg. Springer-Verlag.

Sgall, P. (1967): *Generativní popis jazyka a česká deklinace*. Praha, Academia.

Sgall, P. (1998): Teorie valence a její formální zpracování. *Slovo a slovesnost*. 59, p. 15–29.

Sgall, P. (2006): Valence jako jádro jazykového systému. *Slovo a slovesnost*. 67, p. 163–178.

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

Sgall, P. – Panevová, J. – Hajičová, E. (2004): Deep Syntactic Annotation: Tectogrammatical Representation and Beyond. In Meyers, A. (ed.) *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, p. 32–38, Boston. Association for Computational Linguistics.

Sgall, P. et al. (1969): *A Functional Approach to Syntax in Generative Description of Language*. New York, American Elsevier Publishing Company, Inc. (co-authors: Nebeský, L. – Goralčíková, A. – Hajičová, E.).

Sgall, P. et al. (1986a): *Úvod do syntaxe a sémantiky*. Praha, Academia. (co-authors: Bémová, A. – Borota, J. – Hajičová, E. – Hajičová, I. – Jirků, P. – Panevová, J. – Plátek, M. – Vrbová, J.).

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

Skoumalová, H. – Straňáková-Lopatková, M. – Žabokrtský, Z. (2001): Enhancing the Valency Dictionary of Czech Verbs: Tectogrammatical Annotation. In Matoušek, V. et al. (eds.) *Proceedings of Text, Speech and Dialog International Conference, TSD 2001*, 2166 / *LNAI*, p. 142–149, Berlin Heidelberg. Springer-Verlag.

Straňáková-Lopatková, M. – Žabokrtský, Z. (2002): Valency Dictionary of Czech Verbs: Complex Tectogrammatical Annotation. In González Rodríguez, M. – Paz

Suárez Araujo, C. (eds.) *Proceedings of the 3rd International Conference on Language Resources and Evaluation, LREC 2002*, Vol. 3, p. 949–956, Paris. ELRA.

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

Šindlerová, J. – Bojar, O. (2009): Towards English-Czech Parallel Valency Lexicon via Treebank Examples. In *Proceedings of 8th Treebanks and Linguistic Theories Workshop (TLT)*, p. 185–195, Milano, Italy.

Tesnière, L. (1959): *Eléments de syntaxe structurale.* Paris, Librairie C. Klincksieck.

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 Linguistics*, p. 93–112. Bratislava: Veda.

Urešová, Z. (2010): Building the PDT-VALLEX valency lexicon. In *Proceedings of the Fifth Corpus Linguistics Conference*, Liverpool, UK. (in press).

Uspenskij, V. A. (1977): K ponjatiju diatezy. In Chrakovskij, V. S. (ed.) *Problemy lingvisticheskoj tipologii i struktury jazyka*, p. 65–84. Leningrad.

Wierzbicka, A. (1989): *English Speech Act Verbs: A Semantic Dictionary.* , Academic Press.

Zeman, D. (2004): *Parsing with a Statistical Dependency Model.* PhD thesis, Charles University in Prague, Prague.

Zeman, D. – Sarkar, A. (2000): Learning Verb Subcategorization from Corpora: Counting Frame Subsets. In Gavrilidou, M. et al. (eds.) *Proceedings of the 2nd International Conference on Language Resources and Evaluation, LREC 2000*, Vol. 1, p. 227–233, Athene, Greece.

Zipf, G. K. (1935): *Psycho-Biology of Languages.* Boston, Houghton-Mifflin. (2nd edition MIT Press, 1965).

Žabokrtský, Z. (2005): *Valency Lexicon of Czech Verbs.* PhD thesis, Charles University in Prague, Prague.

Žabokrtský, Z. – Lopatková, M. (2004): Valency Frames of Czech Verbs in VALLEX 1.0. In Meyers, A. (ed.) *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, p. 70–77, Boston. Association for Computational Linguistics.

Chapter A

Building the First Version of the Lexicon

A.1 Enhancing the Valency Dictionary of Czech Verbs: Tectogrammatical Annotation

In Matoušek, V. et al. (eds.) *Proceedings of Text, Speech and Dialog International Conference, TSD 2001, 2166 / LNAI*, p. 142-149, Berlin Heidelberg, 2001. Springer-Verlag
(with co-authors H. Skoumalová and Z. Žabokrtský)

Enhancing the Valency Dictionary of Czech Verbs: Tectogrammatical Annotation

Hana Skoumalová[1], Markéta Straňáková-Lopatková[2], and Zdeněk Žabokrtský[2]

[1] Institute of Theoretical and Computational Linguistics, FF UK, Prague
Hana.Skoumalova@ff.cuni.cz
[2] Center for Computational Linguistics, MFF UK, Prague
{stranak,zabokrtsky}@ufal.mff.cuni.cz

Abstract. A syntactic lexicon of verbs with the subcategorization information is crucial for NLP. Two phases of creating such lexicon are presented. The first phase consists of the automatic preprocessing of source data–particular valency frames are proposed. Where it is possible, the functors are assigned, otherwise the set of possible functors is proposed. In the second phase the proposed valency frames are manually refined.

1 Introduction

In this paper[1] we introduce a semi-automatically prepared syntactic lexicon of Czech verbs that is enriched with information about functors (members of valency frames) on the tectogrammatical (underlying) level of language description (Section 2). Such a lexicon is crucial for any applied task requiring automatic processing of natural language. We focus on verbs because of their central role in the sentence–the information about the modifiers of a particular verb enables us to create the 'skeleton' of the analyzed sentence. It can also be used for example in connection with WordNet for semantic grouping of verbs.

As the source data we use a dictionary of verb frames (originally created at Masaryk University) which is automatically preprocessed (Section 3). In the first phase we only process small set of verbs and their frames. This testing set serves for the estimation of the extent of changes in automatically pre-processed valency frames which must be done manually (Section 4). More extensive sets will follow. We expect that a substantially richer lexicon will be available in several months. In the last section (Section 5) the (preliminary) results are presented.

2 The Concept of Valency Frames of Verbs

Valency theory is a substantial part of the Functional Generative Description of Czech (FGD, [Sgall et al, 1986]), and has been intensively studied since the seventies. Originally it was established for verbs and their frames (see esp. [Panevová, 1974-1975, 1980, 2001]), and was later extended to other parts of speech (nouns and adjectives).

[1] This work has been supported by the Ministry of Education, project LN00A063, and GAČR 405/96/K214.

V. Matoušek et al. (Eds.): TSD 2001, LNAI 2166, pp. 142–149, 2001.

The concept of valency primarily pertains to the level of underlying representation (linguistic meaning) of a sentence and thus it is one of the most important theoretical notions. On the other hand, as the valency information plays a crucial role also for NLP, the morphemic representation of particular members of valency frame is important.

A verbal valency frame (in a strict sense) is formed by so called valency modifiers– that is, the inner participants, either obligatory or optional, together with the obligatory free modifiers. Each Czech verb has at least one valency frame, but it can have more frames. Slots for valency modifiers together with possible morphemic forms of inner participants are stored in a lexicon.

On the level of underlying representation, we distinguish five actants (inner participants) and a great number of free modifiers. The combination of actants is characteristic for a particular verb. Each actant can appear only once in a valency frame (if coordination and apposition are not taken into account). The actants distinguished here are Actor (or Actor/Bearer, Act), Patient (Pat), Addressee (Addr), Origin (Orig) and Effect (Eff). On the contrary, free modifiers (e.g. local, temporal, manner, casual) modify any verb and they can repeat with the same verb (the constraints are semantically based). Most of them are optional and only belong to a 'valency frame' in a broader sense.

The inner participants can be either obligatory (i.e. necessarily presented at the level of underlying representation) or optional. Some of the obligatory participants may be omitted in the surface (morphemic) realization of a sentence if they can be understood as general. Similarly, there exist omissible obligatory free modifiers (as e.g. direction for 'přijít' (to come)). Panevová ([Panevová, 1974-1975]) stated a dialog test as a criterion for the obligatoriness of actants and free modifiers.

FGD has adopted the concept of shifting of 'cognitive roles' in the language patterning ([Panevová, 1974-1975]). Syntactic criteria are used for the identification of Actor and Patient (following the approach of [Tesnière, 1959]), Actor is the first actant, the second is always the Patient. Other inner participants are detected with respect to their semantic roles ([Fillmore, 1968], for Czech [Daneš, Hlavsa, 1981]).

For a particular verb, its inner participants have a (usually unique) morphemic form which must be stored in a lexicon. Free modifiers typically have morphemic forms connected with the semantics of the modifier. For example, a prepositional group Prep 'na' (on) + Accusative case typically expresses Direction, Prep 'v' (in) + Local case has usually local meaning - Where.

In addition to the classical theoretically-based valency also quasi-valency is introduced which may be paraphrased as 'commonly used modification' of a particular item. The concept of quasi-valency enables us to enlarge the information stored in the lexicon, to capture also modifications not belonging to the valency frame in a strict sense ([Straňáková, submitted]). There are free modifiers which are not obligatory (and hence do not belong to the standard valency frame) though they often modify a particular verb. Three sources of such modifiers can be distinguished - (i) 'usual' modifiers without a strictly specified form (like Direction for 'jít' (to go), or Local modifier for 'bydlet' (to stay)), (ii) modifiers with a determined morphemic form (often Regard, e.g., 'zvýhodnit v něčem/na něčem' (to make (st) advantageous for st), or Aim ('potřebovat / poskytovat na něco' (to need / provide (st) for st)), and (iii) theoretically unclear cases with 'wider'

and ‘narrower’ specification (e.g., cause in ‘zemřít na tuberkulózu kvůli nedostatku léků’ (to die of tuberculosis because of the lack of medicine)).

Idiomatic or frozen collocations (where the dependent word is limited either to one lexical unit or to small set of such units, as e.g. ‘mít na mysli’ (to have on mind)) represent specific phenomenon. We resigned on a very complex task of their processing in this stage.

The concept of omissible valency modifiers is reopened with respect to the task of the lexicon. The omissibility of a modifier is not marked in particular lexical entries–we presuppose that in the surface (morphemic) realization of the sentence any member of valency frame is deletable (at least in the specific contexts as e.g. in a question-answer pair).

Analogically, the fact that particular actant can be realized as a general participant is not marked in the valency frame of a verb.

Table 1. Verbal modifiers stored in the lexicon.

	obligatory	optional
inner participants	including general participants	+
free modifiers	including omissible modifiers	“commonly used”

3 Data Preprocessing

As the source data we use a dictionary of verb frames created at Masaryk University ([Pala and Ševeček, 1997], [Horák, 1998]). The lexicon contains valency frames of circa 15,000 Czech verbs. The structure is described in [Horák, 1998].

3.1 Algorithm for Automatic Assigning the Functors

Identifying and merging frames. In the source lexicon, every lemma is listed only once, even if it has several valency frames. A single valency frame, on the other hand, can have several variants (e.g. ‘učit koho co(acc)’, ‘učit koho čemu(dat)’ (to teach sb st)). The variants of one frame are mixed with other frames and thus the first task is to separate the different frames and merge the variants. Let us show it on an example. The verb ‘bránit’ (to protect/prevent) has the following format in the source lexicon:

```
bránit  <v>hTc3,sI,hPc3-sUeN,hPc3-hTc6r{v},hPTc4,
hPTc4-hPTc3r{proti},hPTc4-hPTc7r{před}
```

Single frames are separated by commas and members inside a single frame are separated by dashes. The attribute ‘h’ describes ‘semantic’ features (P-person, T-thing), the attribute ‘c’ stands for morphemic case, ‘r’ means the value of the preposition (in curly braces), ‘sI’ means infinitive and ‘sUeN’ is negative clause with conjunction ‘aby’ (that).

Now, we can arrange the members of all its frames into a table and we can try to find maximal non-intersecting parts.

hTc3			
	sI		
		hPc3 sUeN hPc3 hTc6r {v}	
			hPTc4 hPTc4 hPTc3r{proti} hPTc4 hPTc7r{před}

In the table above we can identify 4 parts. The members that never occur in one frame together can be declared with high probability as variants of one member. Frames with single members (like the first and second frame in the example) can be understood as separate frames, as in the case of 'mířit kam' (to head somewhere), 'mířit na koho' (to aim at sb), or as variants of one frame, as in the case of 'bádat nad čím', 'bádat o čem' (to research into st). We decided to 'merge as much as possible', because of an easier assignment of the functors. The result is shown below.

```
bránit   <v>[hTc3|sI]
bránit   <v>[hPc3]-[sUen|hTc6r{v}]
bránit    <v>[hPTc4]-[hPTc3r{proti}|hPTc7r{před}]
```

Assigning functors. First, we have to add missing subjects to all frames. Then we assign functors to all members of a frame. Unfortunately, there is no straightforward correspondence between the deep frame and its surface realization, but we can try to find some regularities or tendencies, and then formulate rules for assigning the functors to the surface frames. Among all correspondences between the two levels, there are some which are considered as typical. In the direction from the tectogrammatical level to the morphemic one these are:

Actor → Nominative,
Patient → Accusative,
Addressee → (animate) Dative,
Effect → Prep 'na' (to) + Accussative, or Prep 'v' (into) + Accusative,
Origin → Prep 'z' (from) + Genitive, or Prep 'od' (from) + Genitive.

In the opposite direction the correspondences are not so clear because of free modifications, which have a very broad repertory of surface realizations.

For the successful assignment of actants it is necessary to identify free modifiers. The identification is done already during the merging the frames: there exists a list of possible functors for every surface realization, and this list is attached to every member of the original frame. When we merge two members of a frame together we also make an intersection of the attached lists. An empty intersection prevents the two members from being merged. It means that we also get a set of possible functors for every member of a frame as the result of the merging phase. In the optimal case, every member has only one functor assigned.

After identifying free modifiers we can use an algorithm proposed by Panevová and Skoumalová ([Panevová and Skoumalová, 1992]) for the actants. This algorithm is based on the observation that verb frames fall in two categories. The first category contains

frames with at most two actants. The functors are assigned on the base of the 'rule of shifting' (see Section 2)–if there is only one actant in the frame it must be an Actor, and if there are two, one of them is an Actor and the other a Patient. As we had to add subjects automatically, we also made the assumption that they all represent Actor, and thus all frames in this category are already resolved.

The other category contains frames with at least three actants, which can be sorted into two subcategories: prototypical and non-prototypical. The prototypical frames contain only typical surface realizations, and the rule about typical realization can be reverted: if the surface frame contains only typical surface forms we can assign the corresponding functors to them. The non-prototypical frames contain at least one untypical surface realization and a different approach must be adopted. The algorithm is described in [Skoumalová, submitted].

After the merging phase, we get three sorts of frames: frames where every member of a frame has only one functor assigned; the second category contains frames with identified actants but ambiguous free modifiers; and the third category contains frames where at least one member is ambiguous between an actant and a free modifier. Approximately one third of all merged frames (circa 6500) falls into the first category ('final' frames in the sequel) and another thousand into the second category. These frames are candidates for further processing with the help of the above mentioned algorithm, and therefore they will be separated from the rest (circa 11,000), which must be left for manual post-editing (the frames belonging to the second and third category are referred as 'ambiguous'). The editor's work should be easier as s/he gets a (small) set of possible functors which can be assigned to every member of a frame and s/he does not have to choose from all 47 possibilities.

3.2 Testing Set

For the purpose of testing we made a small set containing 178 most frequent Czech verbs with their frames. We omitted the verb 'být' (to be) as it needs a special treatment, and several modal verbs. The set contained circa 350 frames that were created automatically from the source lexicon. They fall into all three categories mentioned above, which means 1) fully resolved frames, 2) frames with ambiguous free modifiers, and 3) frames with ambiguities between actants and free modifiers.

4 Manual Annotation

The data resulting from the preprocessing step are not perfect: they contain incorrectly or ambiguously assigned functors, valency frames proposed may contain mutually excluding (alternating) modifiers, some frames are incorrectly merged into a single one, etc.

That is why we developed a 'tailored' editor for the manual processing of the valency frames of verbs which were pre-processed automatically, as was described above. The editor was implemented as a relational database in Microsoft Access environment.

After obtaining some experiences with annotating the lexicon, we exported the data from the relational database into XML data format (Extensible Markup Language, [Kosek, 2000]). Presently, the XML data are annotated directly in a text editor.

The following attributes are captured for each frame slot:

- functor;
- surface: morphemic realization (mostly morphemic case of a noun or a prepositional group), or a list of possible realization of the particular modifier; the value can be omitted if no special surface realization is required for the given slot (e.g. directional circumstantionals);
- type: this attribute differentiates between obligatory, optional, and quasi-valency modifiers;
- alternative: modifiers, which are mutually excluding, are marked.

4.1 Examples

The following examples illustrate the automatically assigned functors and the manual refinement of valency frames.

The verb 'existovat' (to exist) only has a valency frame that belongs to the first category (fully resolved frames):

existovat R--1[hPTc1]E[hTc2r{u}|hTc6r{na}|hTc6r{v}]$
translated as Actor (Nom) Loc (u+2/na+6/v+6)
manually added mark for arbitrary morphemic realization of local modifier.

The verb 'působit' (to act/operate/work) has been automatically assigned with three valency frames, two of them (1st,3rd) marked as 'ambiguous', one (2nd) as 'final':

působit1 (to operate on st with st) R--1[hPTc1]2CI[hTc7]2A[hPTc4r{na}]& 'ambig.'
translated as Actor (Nom) amb. (na+Acc) amb. (Ins)
manually changed to Actor (Nom) Patient (na+Acc) Means (Ins),
where Actor and Patient are obligatory, Means is a quasi-valency modifier;
působit2 (to do st to sb) R--1[hPTc1]2[hTc4]3[hPc3]& 'final'
translated as Actor (Nom) Patient (Acc) Addr (Dat)
manually the alternative surface forms for Patient are added -
clause attached with conjunctions 'že' (that) or 'aby' (so that);

působit3 (to work as sb)R--1[hPTc1]2P[sU]2JR[hTc4r{jako}]& 'ambig.'
translated as Actor (Nom) amb. (aby) amb. (jako+Acc)
manually changed to Actor (Nom) Patient (jako+Nom) / Loc
where the modifier attached with the conjunction 'aby' belongs to
the second frame (as an alternative representation of Patient),
here the Patient alternates with the Local modifier.

5 Evaluation of Results, Conclusions

In this stage of work only a small testing set of verbs and their frames has been treated. This set serves for clarifying the way of manual processing ('what' and 'how' we want to catch up). After this small lexicon will be brought to perfection it will be used for further

development and testing of automatic precedures. But even on this set of available data some preliminary results can be stated.

It is clear now that even the frames marked as 'final' after the pre-processing must be checked and manually refined—about 35 percent of 'final' valency frames were perfect, i.e. 13 percent from all frames proposed. Fortunately, there was a relatively large number of frames which only 'slightly' differ from the issues wanted–approximately 16 percent of valency frames were correctly merged, but the functors were assigned incorrectly (often 'verba dicendi'), in 20 other percent either one functor is missing in the frame, or is superfluous. About 27 percent of frames were deleted (circa one half as incorrect, one half as frames already detected with other morphemic realization). Then the missing frames were manually added and several cycles of corrections followed. We proceeded a cross checking: we extracted and separately compared sets of frames containing a certain functor, we compared frames of verbs with similar meaning etc.

Basic statistical characteristics are presented:

- number of the processed verbs: 178
- number of the frames: 462 (in average 2.6 frames per a verb)
- number of all frame slots: 1481 (in average 3.2 slots per a frame)
- distribution of the number of frame slots per a frame (Table 2)
- distribution of frame slots according to their type (Table 3)
- number of occurences of individual functors in the lexicon (Table 4).

Table 2. Distribution of the number of frame slots per a frame.

number of slots	1	2	3	4	5	6	7	8
number of frames	16	145	134	95	45	15	10	1
% (out of all frames)	3.5	31.4	29.0	20.6	9.7	3.2	2.2	0.2

Table 3. Distribution of the frame slots according to the type.

type	obligatory	optional	quasi-valency
occurences	918	200	363
% (out of all slots)	62.0	13.5	24.5

Table 4. Number of occurences of 18 most frequent functors.

order	functor	occurences	order	functor	occurences
1	ACT (actor)	460	10	ORIG (origin)	40
2	PAT (patient)	362	11	DIR1 (direction to)	25
3	ADDR (addressee)	93	12	BEN (benefactive)	23
4	EFF (effect)	86	13	AIM (aim)	21
5	MANN (manner)	71	14	ACMP (accompaniment)	18
6	REG (regard)	67	15	TWHEN (time-when)	15
7	LOC (location)	49	16	DIR2 (dir. which way)	14
8	DIR3 (direction from)	49	17	EXT (extent)	13
9	MEANS (means)	48	18	INTT (intention)	7

Roughly one half of the processed verbs is contained in the Czech part of EuroWordNet lexical database [Pala, Ševeček, 1997]. Currently we try to map the valency frames to EuroWordNet synsets.

We expect that the large amount of time consumed by the preparation of such a small lexicon has its source in the fact that we have processed the most frequent Czech verbs, which likely belong to the most difficult ones. The extension of data processed may lead (and we hope so) to an increased effectiveness of the algorithm presented.

References

1. Daneš, Fr., Hlavsa, Z.: Větné vzorce v češtině. Academia, Praha, 1981.
2. Fillmore, C.J.: The Case for Case. In: Universals in Linguistic Theory (eds. E. Bach, R. Harms), New York, 1-90, 1968.
3. Horák, A.: Verb valency and semantic classification of verbs. In: TSD'98 Proceedings (eds. Sojka, P., Matoušek, V., Pala, K., and Kopeček, I.), Masaryk University Press, Brno, pp.61-66, 1998.
4. Kosek, J.: XML pro každého. Grada Publishing, Prague, 2000.
5. Pala, K., Ševeček, P.: Valence českých sloves (Valency of Czech verbs). In: Sborník prací FFBU, volume A45, Masaryk University, Brno., pp. 41-54, 1997.
6. Pala, K., Ševeček, P.: Final Report, June 1999, Final CD ROM on EWN1,2,LE4-8328, Amsterdam, September 1999.
7. Panevová, J: On Verbal Frames in Functional Generative Description. Part I, PBML 22, 1974, pp.3-40, Part II, PBML 23, pp. 17-52, 1975.
8. Panevová, J.: Formy a funkce ve stavbě české věty. Academia, Praha, 1980.
9. Panevová, J: Valency Frames: Extension and Re-examination, 2001. (submitted)
10. Panevová, J., Skoumalová, H.: Surface and deep cases. In: Proceedings of COLING '92, Nantes, pp. 885-889, 1992.
11. Sgall, P., Hajičová, E., Panevová, J.: The Meaning of the Sentence in Its Semantic and Pragmatic Aspects (ed. by J. Mey), Dordrecht:Reidel and Prague:Academia, 1986.
12. Skoumalová, H.: Czech syntactic lexicon. PhD thesis, Charles University, Faculty of Arts, Prague. (submitted)
13. Straňáková, M.: Homonymie předložkových skupin a možnost jejího automatického zpracování, PhD thesis, MFF UK (submitted).
14. Tesnière, L.: Elements de syntaxe structurale. Paris, 1959.
15. Žabokrtský, Z.: Automatic Functor Assignment in the Prague Dependency Treebank. In: TSD2000 Proceedings, LNAI 1906, Springer-Verlag, pp. 45-50, 2000.

A.2 Valency Dictionary of Czech Verbs: Complex Tectogrammatical Annotation

In González Rodríguez, M. – Paz Suárez Araujo, C. (eds.) *Proceedings of the 3rd International Conference on Language Resources and Evaluation, LREC 2002*, Vol. 3., p. 949-956, Paris, 2002. ELRA
(with co-author Z. Žabokrtský)

Valency Dictionary of Czech Verbs: Complex Tectogrammatical Annotation

Markéta Straňáková-Lopatková, Zdeněk Žabokrtský

Center for Computational Linguistics
Faculty of Mathematics and Physics, Charles University
Malostranské nám. 25, CZ-11800 Prague, Czech Republic
{stranak,zabokrtsky}@ckl.mff cuni.cz
http://ckl.mff.cuni.cz

Abstract

A lexicon containing a certain kind of syntactic information about verbs is one of the crucial prerequisities for most tasks in Natural Language Processing. The goal of the project described in the paper is to create a human- and machine-readable lexicon capturing in detail valency behavior of hundreds most frequent Czech verbs. Manual annotation effort consumed at this project limits the speed of its growth on the one hand, but guarantees significantly higher data consistency than that of automatically acquired lexicons. In this paper, we outline the theoretical background on which the lexicon is based, and describe the annotation schema (lexicon data structure, annotation tools, etc.). Selected quantitative characteristics of the lexicon are presented as well.

1. Introduction

The verb is traditionally considered to be the center of the sentence, and thus the description of syntactic behavior of verbs is a substantial task for linguists. A syntactic lexicon of verbs with the subcategorization information is obviously crucial also for many tasks in Natural Language Processing (NLP) domain. We briefly exemplify the potential contribution of the valency lexicon to several well-known tasks in NLP:

- **Lemmatisation** (choosing the correct lemma for each word in a running text). Example sentences:

  (1) *Stali se matematiky.*
  [They become mathematicians.]

  (2) *Báli se matematiky.*
  [They were afraid of mathematics.]

  In both sentences, the word form *matematiky* occurs. It could be either Acc.pl or Instr.pl of the lemma *matematik* [mathematician] or Gen.sg, Nom.pl, Acc.pl of lemma *matematika* [mathematics]. The lemma can be disambiguated in both sentences using the fact that the verb *stát se* [to become] (sentence 1) contains[1] neither Gen nor Acc in its valency frame, and no frame of the verb *bát se* [to be afraid] (sentence 2) contains Acc or Instr.[2]

- **Tagging** (choosing the correct morphological tag for the given word and lemma). Example:

  (3) *Ptala se jeho bratra.*
  [She asked his brother.]

  The noun phrase *jeho bratra* [his brother] preceded by no preposition can be Gen.sg or Acc.sg. The verb *ptát se* [to ask] allows only the former possibility.

- **Syntactic analysis** (considering a dependency oriented formalism, syntactic analysis can be informally expressed as 'determining which word depends on which'). Examples:

  (4) *Nechala ho spát.*
  'she let him to sleep'
  [She let him sleep.]

  (5) *Začala ho milovat.*
  'she started him to love'
  [She started to love him.]

  In sentence 4 the pronoun *ho* [him] (Gen.sg, Acc.sg) can depend only on the preceding verb *nechat* [to let] (since this verb has a valency frame containing both Acc and infinitive, whereas the valency frame of *spát* [to sleep] contains neither Gen nor Acc). On the other hand, in sentence 5 the same pronoun must depend on the following verb (since no frame of *začít* [to begin] contains both accusative and infinitive). Considering only the morphological tags of the words, both sentences are equivalent. An unambiguous dependency structure[3] cannot be constructed without considering valency frames of the respective verbs.

- **Word sense disambiguation**. Examples:

  (6) *Odpovídal na otázky.*
  [He was answering questions.]

  (7) *Odpovídal za děti.*
  [He was responsible for children.]

  (8) *Odpovídal popisu.*
  [He matched the description.]

[1] In this context, we use 'frame X contains Y' to express the fact that some element of the valency frame X is prototypically realized by the form Y (direct or prepositional case, etc.) on the surface.

[2] The possibility of Nom is excluded in both sentences according to the subject-verb agreement.

[3] A similar claim holds for phrase structure of given sentences.

Different meanings of the same word are often indicated by a change in the valency frames. The meaning of verb *odpovídat* in sentence 6 is 'to answer', in sentence 7 the same word expresses 'to be responsible', and in sentence 8 it expresses 'to match'.

- **'Semantic analysis'**. Examples:

  (9) *Přišel po Petrovi.*
  He came after Peter.

  (10) *Sháněl se po Petrovi.*
  [He seeked for Peter.]

  Prepositional groups most frequently represent adjuncts (as in sentence 9); however, they can also stand for verbal participants (as in 10), which is a crucial difference in most semantically or logically motivated approaches. The role of the prepositional group *po Petrovi* [after / for Peter] cannot be determined without considering valency frames of the respective verbs.

- **Machine translation.** All of the problems mentioned above inevitably arise during any serious attempt at machine translation (MT). Since the existence of a valency dictionary would lead to a higher quality of the respective submodules of such an MT system, it should also increase the quality of the resulting translation.

Existing lexicons for Czech (see Section 4) either do not contain information needed for automatic syntactic analysis, or their coverage is strictly limited, or they are not available in an electronic form, or they are not sufficiently reliable. The consistency is a great problem for most of them.

We present a lexicon of Czech verbs containing rich syntactic information, where the valency information is the most important one. A great emphasis is laid on the formulation of precise criteria for setting the valency frames of particular verbs and their properties, which seems to be a necessary condition for a consistent treatment of the considered phenomena. The lexicon items refer (through Czech WordNet) to EuroWordNet (EWN), which increases the usability of the lexicon for NLP. Emphasis is laid also on both human- and machine-readability of the resulting lexicon.

2. Theoretical Background

2.1. Functional Generative Description

Valency theory is a substantial part of the Functional Generative Description, FGD (Sgall et al., 1986), a dependency oriented description that serves as our theoretical framework. Valency of verbs has been intensively studied since the seventies (Panevová, 1974-75; Panevová, 1980; Panevová, 2001). The concept of valency primarily pertains to the level of underlying representation of a sentence (i.e. the level of linguistic meaning, in FGD called tectogrammatical level). For NLP, also morphemic representation of particular members of the valency frame is important.

The lexical entry for a verb enumerates valency frame(s), at least one but usually more. A valency frame of a verb (in a broader sense) is interpreted as a range of syntactic elements (verbal modifiers) either required or specifically permitted by this verb. It describes a verb in its primary as well as secondary, 'shifted' use (e.g. *tlačit na někoho* [to urge sb / to press on sb]).

The **valency frame** (in a strict sense) of a particular verb consists of valency slots corresponding to inner participants, i.e. actants (both obligatory and optional), and obligatory modifiers (adjuncts, see below).

On the level of underlying representation, we distinguish five **actants** (inner participants) and a wide scale of modifiers. The actants satisfy the following two conditions:

- The combination of actants is characteristic for a particular verb.
- Each actant can appear only once within any occurrence of a particular verb (if coordination and apposition are not taken into account).

The actants distinguished in FGD are Actor (or Actor/Bearer, Act), Patient (Pat), Addressee (Addr), Origin (Orig) and Effect (Eff). Some typical illustrative examples below are taken from the studies of Panevová (quoted in the References).

(11) *Matka*.Act *předělala dětem*.Addr *loutku*.Pat *z Kašpárka*.Orig *na čerta*.Eff.
[Mother.Act re-made a puppet.Pat for children.Addr from a Punch.Orig to a devil.Eff.]

On the contrary, modifiers (e.g. local, temporal, manner, causal) can modify any verb and they can occur repeatedly with the same verb (the constraints are semantically based) - therefore we call them **free modifiers**. Most of them are optional and belong to the 'valency frame' only in a broader sense (for the list of free modifiers see e.g. (Hajičová et al., 2000)). Examples:

(12) *V Praze*.Loc *se sejdeme na Hlavním nádraží*.Loc *u pokladen*.Loc.
[In Prague we will meet at the Main Station near the booking-offices.]

(13) *Kvůli dešti*.Caus *musel čekat pod střechou, protože neměl deštník*.Caus.
'because of rain (he) had to wait under the roof because he didn't have an umbrella'
[As it was raining he had to wait under the roof because he didn't have an umbrella.]

The inner participants can be either **obligatory** (i.e. necessarily present at the level of the underlying representation) or **optional**. Panevová (1974-75) formulated a **dialogue test** as a criterion for the obligatoriness of actants and free modifiers. Informally, the obligatoriness of a modifier means that both the speaker and the listener must know the information expressed by this modifier.[4]

[4]Some of the obligatory participants may be omitted in the surface (morphemic) realization of a sentence, e.g., Actor can be omitted in every Czech sentence. Similarly, free modifiers (both obligatory and optional) are omittable in the surface realization (as e.g. direction for *přij´ı t* [to come], which always means *přij´ı t někam* [to come somewhere]). For the smoothness of the dialogue, both the speaker and the listener must know the necessary information (e.g. from the preceding dialogue or from the broader situation).

	obligatory	optional
inner participants	+	+
free modifiers	+	–

Figure 1: Valency slots creating verbal valency frame (in a strict sense) are marked with '+'

FGD has adopted the concept of **shifting of 'cognitive roles'** in the language patterning (Panevová, 1974-75). Syntactic criteria are used for the identification of Actor and Patient (following the approach of (Tesnière, 1959)), Actor is the first actant, the second is always the Patient. Other inner participants are detected with respect to their semantics

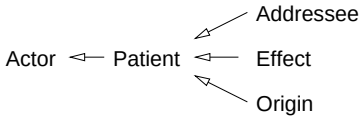


Figure 2: Shifting of cognitive roles.

In other words, if a particular verb has a single actant, it is the Actor (ex. (14)), a verb with two actants has Actor and Patient (regardless the semantics, ex. (15)). The semantics is taken into account with the third and further actants. Examples:

(14) *Škola*.Act *začala.*
[The school lessons began.]

(15) *Bavlně*.Pat *se nic*.Act *nevyrovná.*
[Nothing is as good as cotton.]

(16) *Chlapec*.Act *vyrostl v muže*.Pat
[A boy grew up to a man.]

(17) *Z vašich slov*.Pat *plyne, že zítra nepřijdete*.Act
[It follows from your words that you will not come tomorrow.]

2.2. Enriched Valency Frames

The 'standard' valency view applied in FGD is enriched for the purposes of automatic processing here. In addition to the valency slots creating the valency frame in a strict sense (which does not contain optional free modifiers) also quasi-valency and typical modifiers are stored in the lexicon.

Quasi-valency modifiers are free modifiers that are not obligatory, although they often modify particular verbs and they may specify their meaning (primary, secondary or idiomatic). They can be characterized as 'commonly used modifiers'.

Three sources of quasi-valency modifiers can be distinguished:

- 'usual' modifiers without a strictly specified form (e.g. Direction for verbs of motion, like *jít* [to go]),
- modifiers with a determined morphemic form (e.g. Means in *hrát na kytaru* [play the guitar]), and
- cases with a competition of two occurrences of the modifier, a 'narrower' and a 'wider' specification; the former one is understood as a quasi-valency modifier (e.g. Cause in *zemřít na tuberkulózu kvůli nedostatku léků* [to die of tuberculosis because of the lack of drugs]).

The introduction of **typical modifiers** allows to save all information from the source lexicons. They do not specify the meaning of the verb but they are typical for whole sets of verbs. They usually have a typical form (e.g. Instrumental case for Means as in *psát tužkou* [to write with a pencil], *jet vlakem* [to go by train], or the prepositional group *pro* [for] + Acc for Benefactive as in *pracovat pro firmu* [to work for firm]). In addition, they enable us to capture other syntactic phenomena, such as reciprocity etc. (as described in section 3).

We refer to valency frames capturing valency slots (actants and obligatory free modifiers) as well as quasi-valency and typical modifiers as to **enriched valency frames**.

	obligatory	optional
inner participants	+	+
free modifiers	+	quasi+typical

Figure 3: Modifiers captured in enriched valency frame

For a particular verb, its inner participants have a (usually unique) **morphemic form**, which must be stored in a lexicon (though a prototypical expression of each actant exists, as Nom case for Actor and Acc case for Patient in active sentence, or Dat for Addressee). Free modifiers typically have several different morphemic forms related to the semantics of the modifier. For example, a prepositional group *na* [on] + Acc typically expresses Direction, Prep *v* [in] + Loc has usually local meaning - Where.

The concept of **omissible valency modifiers** is reopened with respect to the task of the lexicon. In principle, conditions of omissibility of particular valency slots on the surface are not yet formally described. We assume that any valency slot is deletable (at least in the specific contexts as e.g. in a question-answer pair).

3. Structure of the Lexicon

3.1. What should a dictionary ideally capture?

The idea is to create lexicon containing all syntactic information useful for NLP. The model proposed offers a complex information on the lexical item (verb), information on its valency frames as well as information specifying elements of these frames.

There is a list of enriched valency frames for each verb (each verb has at least one valency frame, but it may have more frames, with respect to the number of its meanings; primary, secondary as well as idiomatic usage is taken into account).

Several attributes are specified for each valency frame: an ordered sequence of valency slots, a specification of the lexical meaning, examples of usage, the aspectual counterpart, lemma, types of possible diatheses, and pointer(s) to

EuroWordNet synset(s) are the most important ones (see below).

Each frame slot is characterized by a 'functor' (name of an inner participant or modifier, see (Žabokrtsky et al., 2002)), by the type of relation (obligatory, optional and 'quasi-valency' or 'typical' modifier) and by its possible morphemic realization(s).

3.2. Information included in an enriched valency frame

Valency slots. We take over all principles described in section 2. Slots representing valency modifiers are ordered in systemic ordering (introduced in (Sgall et al., 1986)), which reflects unmarked word order in Czech sentence.

Synonyms and examples. A set of synonyms or 'nearly synonyms' together with example(s) of usage specify a particular meaning of the verb.

Alternative frames. A number of verbs exists where a unique meaning can be expressed by two sets of modifiers (e.g. obligatory Addressee and Direction-where often alternates as in *poslal dárky dětem* [he sent gifts to children] / *poslal dárky do Konga* [he sent gifts to Congo]). Such valency frames are marked as alternative frames.

Reciprocity. A concept of reciprocity (Panevová, 1999) expresses the possibility of some modifiers of the given verb to be symmetrical (as in a sentence *Jan a Marie se milují* [John and Mary are in love] where both members *Jan* and *Marie* can be interpreted as Actor and Patient). The possibility of reciprocal use of a verb (in its particular sense) is marked in the lexicon - for relevant valency frames there is a list of modifiers that can be in the relation of reciprocity.

Control. Generally, the notion of control relates to a certain type of predicate (verb of control) and two coreferential expressions, a controller and a controllee. We focus on a situation where a verb has an infinitive modifier (regardless its functor). Then controllee is the member that would be the 'subject' of infinitive (which is structurally excluded on the surface), controller is the co-indexed member of the particular valency frame of the head verb (Panevová, 1997); the controller is marked in the lexicon, see also (Skoumalová, 2001). (E.g. the verb *pokoušet se* [to attempt at st] has Patient which can be expressed by an infinitive; its Actor is marked as the controller - see sentence *Marie se pokouší zpívat* [Mary attempts at singing] where *Marie* being the Actor of the head verb *pokoušet se* is the 'subject' of the dependent verb *zpívat* [to sing].)

Diathesis. The lexicon contains valency frames for the active voice of verbs. Many of the diatheses, especially passive constructions are derived regularly (Skoumalová, 2001), thus the individual valency frames are marked only with a marker showing which types of diatheses can be derived from the active form. Only the exceptions are treated explicitly.

Aspectual counterparts. Usually, lexicons designed for human readers list lexical items only for imperfect verbs (which are considered to be the primary ones). The lexicon described here contains separate lexical items for both aspects of verb, the aspectual counterparts are connected with pointers. There are two reasons for this decision:

*** bránit** [to defend / to restrain / to obstruct]
-aspect: *(imp.)*
+ ACT(1;obl) PAT(4;obl) EFF(před+7,proti+3;obl)
-synon: *zajišťovat obranu*
-example: *Obyvatel´e br´an´ı město před Švédy, před ´utoky.*
[The inhabitants defend a town against the Swedes, against attacks.]
-use: *prim*
-freq: *3*
-ewn: *2*
+ ACT(1;obl) ADDR(3;obl) PAT(v+6,Inf,aby;obl) MEANS (7;typ)
-synon: *zabraňovat, držst zp´atky*
-example: *Br´an´ı mu v tom všemi silami.*
[He impedes him in it with all means.]
-reciprocity: *ACT-ADDR*
-control: *ADDR*
-use: *posun*
-freq: *15*
-ewn: *1*
+ ACT(1;obl) PAT(3;obl) MEANS(7;typ)
-synon: *zabraňovat*
-example: *Petr br´anit jejich štěst´ı .*
[Peter obstructs their happiness.]
-use: *posun*
-ewn: *1*
*** bránit se** [to prevent]
-aspect: *(imp.)*
+ ACT(1;obl) PAT(3,proti+3,před+7;opt) MEANS(7;typ)
-synon: *chr´anit se*
-example: *Br´an´ı se vyd´ı r´an´ı ; proti vyd´ı r´an´ı .*
[They prevent themselves against a blackmail.]
-use: *prim*
-freq: *7*

Figure 4: A sample from the valency lexicon

- generally, valency frames may differ for perfect and imperfect aspect of a verb, especially for its secondary or idiomatic usage, and
- the aspectual pairs are treated separately in the Czech WordNet, and thus the pointers to EWN differ for these pairs.

Primary / secondary / idiomatic usage. The valency frames of a particular verb are ordered according to the type of usage - we distinguish primary, secondary and idiomatic usage. This ordering (generally more or less corresponding to the frequency of particular frames - tested on a sample of Czech National Corpus, CNC, (Čermák, 2001)) contribute to an easier orientation in the lexicon. In this stage of work, idiomatic or frozen collocations (where the dependent word is limited either to one lexical unit or to small set of such units, as e.g. *mít na mysli* [to have on mind]) is only partially treated.

Syntactic/semantic classes. Though different semantic classifications of verbs exist, none of them seems to be really appropriate for our task. We preliminarily classify the verbs into several syntactic/semantic classes, such as

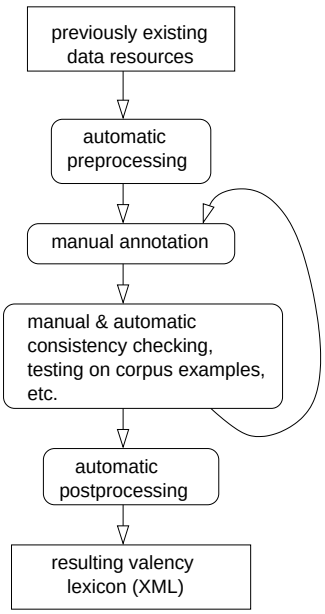


Figure 5: Data flow diagram.

verba dicendi, verbs of movement or verbs of exchange, etc. Such classification helps us when checking the lexicon consistency (verbs from the same class should be treated similarly).

Pointers to Czech WordNet. Valency frames of verbs from the lexicon contained also in Czech WordNet (Pala, Ševeček, 1999) have a pointer to the corresponding Czech synset(s) (=set of synonyms) and through it/them to the interlingual semantic database EuroWordNet (see http://www.hum.uva.nl/ ewn/).

4. How is the Lexicon Created

4.1. Data Resources

Dictionary of verb frames. When creating the lexicon, we utilize other existing electronic resources for Czech. First of all, it is the dictionary of verb frames built up at the Masaryk University (Pala, Ševeček, 1997). The lexicon contains possible morphemic realizations of valency frames of ca 15 000 Czech verbs. Its structure is described in (Horák, 1998). This machine-readable lexicon does not contain information about underlying 'functors' of particular valency frames, the particular meanings of verbs are not specified.[5]

Slovesa pro praxi (Verbs for practise, (Svozilová et al., 1997)). This valency lexicon containing a detailed analysis of ca 750 frequent Czech verbs offers substantial information. Unfortunately, its coverage is limited and the conception of this manually processed lexicon excluded automatic exploitation.

Prague Dependency Treebank. The processing of verbs is based on a number of analyses in theoretical articles concerning FGD, especially those of Panevová. Many unclear aspects are discussed during tectogrammatical annotation of the Prague Dependency Treebank, PDT (Hajičová et al., 2000).

Czech National Corpus. We intensively use the Czech National Corpus, CNC (Čermák, 2001), which serves especially for the verification of valency frames stated and for filling in the gaps.

EuroWordNet and Czech WordNet. The semantic database EuroWordNet (see http://www.hum.uva.nl/ ewn/) and especially its Czech part (Pala, Ševeček, 1999) with its conception of synsets (sets of synonyms, or 'nearly synonyms') contributes to the specification of particular verb meanings.

Slovn´ıkčeské frazeologie a idiomatiky (Lexicon of Czech Phraseology and Idioms, (Čermák, Hronek, 1983)). Though our approach is much more syntax-based, the lexicon of idiomatic expressions helps with the treatment of idioms.

4.2. Annotation

There have been several attempts at creating a valency lexicon automatically but the output of such efforts is not satisfactory. Unfortunately, the great extent of manual annotation seems to be unavoidable for this task, but existing resources can be used which makes it more effective (namely WordNet for Czech, dictionary of morphemic characterization of modifiers of particular verbs, syntactically and morphologically tagged corpora and others).

The lexicon arises in batches of roughly 100 verbs (according to the frequency in the PDT). The 'coverage' of the individual batches is depicted in Figure 6. The process is divided into two steps: automatic preprocessing and manual annotation. In the first step, the resources available are added to all verbs and a preliminary functor assignment is curried on. The second step consists mainly of splitting and merging frames, assigning the functors and correcting the automatically prepared ones, adding the examples. Mapping particular frames on EuroWordNet synset(s) is another important task of the human annotator.

4.3. Software Tools, Data Representation

In order to make the manual annotation as fast as possible, comfortable and effective tools must have been created.

The main annotation tool is the annotation editor. Currently we use a customizable text editor WinEdt (see Figure 7) with a special mode tailored for our lexicon. The data are represented as a (structured) plain text: each line starting with '*' contains a lemma, each line starting with '+' contains a valency frame (written as a sequence of functors followed by parentheses containing surface realization and type of the slot), each line starting with '-' contains a frame attribute (attribute name followed by ':' and attribute value). A (simplified) sample of the data is given in Figure 4.

This approach allows an extremely easy manipulation with lexicon data structures and brings no overhead operations for the annotator. Since the mode colorizes the lexicon data (syntax highlighting), the navigation is also very comfortable.

The second most important tool is the search engine that allows to search for valency frames (in the already ex-

[5]Let us notice also **valency lexicon** that has been **automatically created** on the basis of this dictionary, see (Skoumalov´a, 2001).

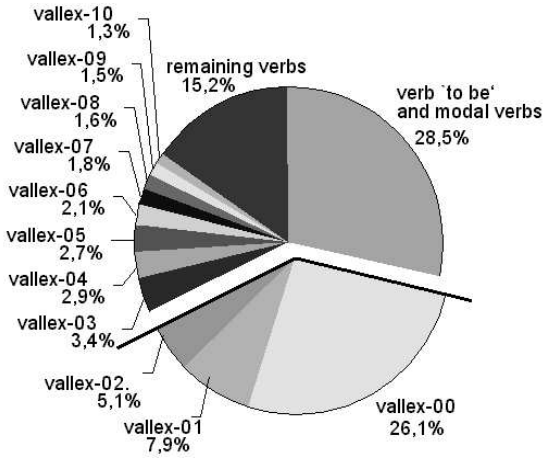


Figure 6: 'Coverage' of the lexicon tested on the verbs in running text from the Czech National Corpus. Vallex-00 contains roughly 160 verbs, each of the remaining batches contains roughly 100 verbs each. The thick line picks out the portion of verbs the annotation of which has been practically finished.

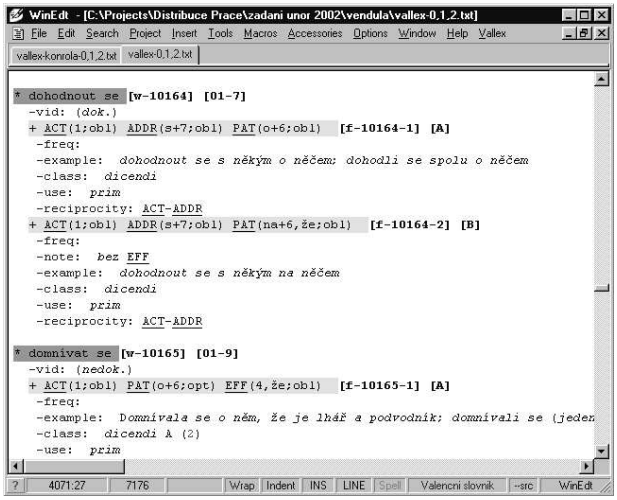


Figure 7: WinEdt screenshot.

isting part of the lexicon) according to a specified query. For example, those frames can be automatically searched which were classified as verba dicendi, have adressee slot expressed by dative.

4.4. Verification, Cross-Checking

We lay a great emphasis on the consistency of the lexicon. The completeness of the data is checked in comparison with the CNC (for each verb a set of sentences is chosen and the annotators 'maps' the occurrences of the verb onto particular valency frames; if need, new frame(s) are added).

The software tools developed allow for sorting valency frames according to a scale of attributes (verb class, morphemic form of modifiers, presence of particular valency slot etc.), which contributes to a consistent treatment of particular phenomena (let us mention e.g. a sometimes unclear boundary between Addressee and Benefactive, or systematic processing of verbs belonging to one class).

The lexicon is used for (manual) tectogrammatical annotation of the PDT. It means a systematic practical verification of the concept accepted as well as of the completeness of the data.

4.5. Selected quantitative characteristics of the data

The project reported on is in progress. The first set of ca 160 verbs served for the development and verification of the annotation scheme, the methodology and the software tools.

At present, a set of 331 most frequent verbs is processed (and used by PDT annotators), as is shown in Figure 6. There are 1110 valency frames for these verbs, which contain altogether 3317 valency slots. Various statistical characteristics are given in Figures 8, 9, and 10.

Another set of 200 verbs is almost completed. Modal verbs and auxiliary *být* [to be], which have been excluded in the first stages as they need a special treatment, is processed now.

We assume that another set of ca 600 verbs will be completed till summer 2002 (it means a 'coverage' of about 85% on the verbs in running text from CNC, see 'remaining verbs' in Figure 6].

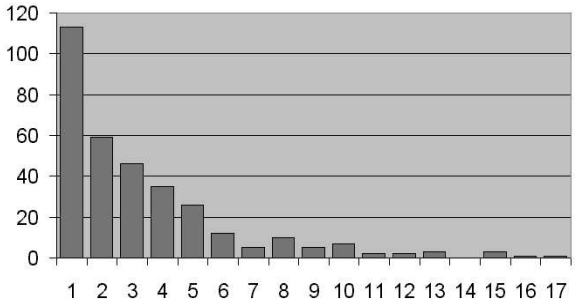


Figure 8: Distribution of the number of valency frames per a lemma.

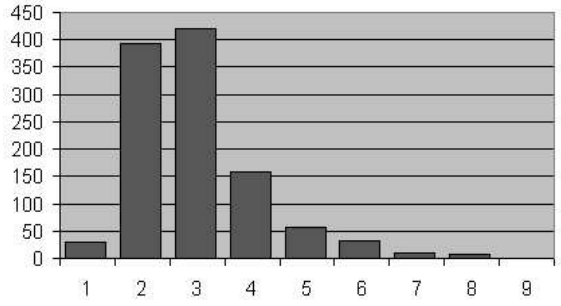


Figure 9: Distribution of the number of valency slots per a frame.

5. Closing remarks

5.1. Open problems

A systematic processing of verbs asks for clear (syntactically based) principles of annotation. Till now, several important questions remain open; though some of them are entirely theoretically described we still miss reliable criteria. The following problems are the most relevant:

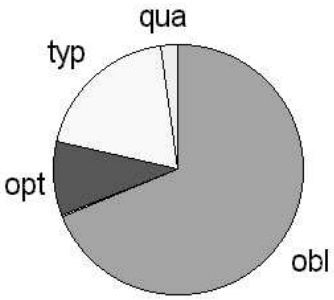


Figure 10: Distribution of values of the type of frame slots.

- The difference between a concrete and an abstract meaning of a verb (e.g. Direction for *vycházet z lesa* [to leave a forest] vs. Direction / Patient for *vycházet z předpokladů* [to start from the premises]).
- Criteria for the distinguishing particular verb meanings (too coarse-grained 'pure syntactic' criteria vs. too fine-grained classification of EWN).
- Criteria for the determination whether a verb with the reflexive particle *se / si* [6] constitutes a separate lexical unit. Example:

  (18) *Matka myje dítě houbou.*
  [Mother washes a child with a sponge.]

  (19) *Myji se každé ráno studenou vodou.*
  [I wash myself every morning with cold water]

  These two Czech sentences exhibit the same syntactic structure; nevertheless, the verbs *mýt* and *mýt se* can be treated in some approaches as two units.
- A complex treatment of idioms.

5.2. Conclusion

We have presented the concept of the lexicon of Czech verbs containing all syntactic phenomena which may be useful for NLP. Though some questions remain open in this stage of our work, the sample of the lexicon (containing 331 most frequent verbs) is successfully used in the process of annotating PDT. A substantial extension is presupposed before summer 2002.

We have mentioned the tasks in NLP to which the lexicon can contribute. On the other hand, it can be useful also for a theoretically based research - the lexicon can be used e.g. for capturing valency of other word classes.

Acknowledgement

The research reported on in this paper has been carried out under the projects MŠMT LN00A063.

6. References

Fr. Čermák. 2001. Language Corpora: The Czech Case. In: V. Matoušek, P. Mautner, R. Mouček, K. Taušer (eds), *Proceedings of the 4th International Conference on Text, Speech and Dialogue - TSD2001*. LNAI 2166. Springer. 21-30.

Fr. Čermák, J. Hronek. 1983. Slovník české frazeologie a idiomatiky (Lexicon of Czech Phraseology and Idioms). Praha. ČSAV.

Fr. Daneš, Z. Hlavsa. 1981. Větné vzorce v češtině (Sentence Patterns in Czech). Praha. Academia.

Ch. J. Fillmore. 1968. The Case for Case. In: E. Bach, R. Harms (editors), *Universals in Linguistic Theory*. New York. 1-90.

E. Hajičová, J. Panevová, P. Sgall. 2000. A Manual for Tectogrammatical Tagging of the Prague Dependency Treebank. UFAL/CKL Technical Report TR-2000-09.

A. Horák. 1998. Verb valency and semantic classification of verbs. In: P. Sojka, V. Matoušek,K. Pala, I. Kopeček, (eds.), *Proceedings of the First Workshop on Text, Speech, Dialogue - TSD'98*. Brno. Masaryk University Press. 61-66.

K. Pala, P. Ševeček. 1997. Valence českých sloves (Valency of Czech verbs). In: *Sborník prací FFBU*. volume A45. Brno. Masaryk University. 41-54.

K. Pala, P. Ševeček. 1999. Final Report Brno, June 1999, Final CD ROM on EWN1,2,LE4-8328. Amsterdam. September 1999.

J. Panevová. 1974-75. On Verbal Frames in Functional Generative Description. Part I. PBML 22. 3-40. Part II. PBML 23. 17-52.

J. Panevová. 1980. Formy a funkce ve stavbě české věty (Forms and Functions in the syntax of Czech sentence). Praha. Academia.

J. Panevová. 1994. Valency Frames and the Meaning of the Sentence. In: P. A. Luelsdorff (ed.), *The Prague School of Structural and Functional Linguistics.* Amsterdam, Philadelphia. Benjamins Publ. Comp. 223-243.

J. Panevová. 1997. More Remarks on Control. In: E. Hajičová, O. Leška, P. Sgall, Z. Skoumalová (eds.), *Prague Linguistic Circle Papers*. Vol. 2. Amsterdam-Philadelphia:John Benjamins. 101-120.

J. Panevová. 1999. Česká recipročná zájmena a slovesná valence (Czech reciprocy pronouns and valency of verbs). Slovo a slovesnost 60. 269-275.

J. Panevová. 2001. Valency Frames: Extension and Reexamination. In: V. S. Charkovskij, M. Grochowski, G. Hentschel (eds.), *Festschrift fuer Andrzej Boguslawski* Studia Slavica Oldenburgensia. No. 9. Oldenburg. Bibliotheks- und Informationssystem. 325-340.

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

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

H. Skoumalová, Straňáková-Lopatková, Žabokrtský. 2001. Enhancing the Valency Dictionary of Czech Verbs: Tectogrammatical Annotation. In: V. Matoušek, P. Mautner, R. Mouček, K. Taušer (eds), *Proceedings of the 4th International Conference on Text, Speech and Dialogue - TSD2001*. LNAI 2166. Springer. 142-149.

N. Svozilová H. Prouzová, A. Jirsová. 1997. Slovesa pro praxi (Verbs for Practice). Praha. Academia.

[6] Now reflexive passive and reciprocity are not taken into account.

L. Tesnière. 1959. Elements de syntaxe structurale. Paris.
Z. Žabokrtský, P. Sgall, S. Džeroski. 2002. A Machine Learning Approach to Automatic Functor Assignment in the Prague Dependency Treebank. In *Proceedings of LREC 2002*.

A.3 Valency Lexicon of Czech Verbs VALLEX: Recent Experiments with Frame Disambiguation

In Matoušek, V. – Mautner, P. – Pavelka, T. (eds.) *Proceedings of Text, Speech and Dialog International Conference, TSD 2005, 3658 / LNAI*, p. 99-106, Berlin Heidelberg, 2005. Springer-Verlag
(with co-authors O. Bojar, J. Semecký, V. Benešová and Z. Žabokrtský)

Valency Lexicon of Czech Verbs VALLEX: Recent Experiments with Frame Disambiguation*

Markéta Lopatková, Ondřej Bojar, Jiří Semecký,
Václava Benešová, and Zdeněk Žabokrtský

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics, Charles University, Prague
{lopatkova,bojar,semecky,benesova,zabokrtsky}@ufal.mff.cuni.cz

Abstract. VALLEX is a linguistically annotated lexicon aiming at a description of syntactic information which is supposed to be useful for NLP. The lexicon contains roughly 2500 manually annotated Czech verbs with over 6000 valency frames (summer 2005). In this paper we introduce VALLEX and describe an experiment where VALLEX frames were assigned to 10,000 corpus instances of 100 Czech verbs – the pairwise inter-annotator agreement reaches 75%. The part of the data where three human annotators agreed were used for an automatic word sense disambiguation task, in which we achieved the precision of 78.5%.

1 Introduction

A verb is traditionally considered to be the center of the sentence, and description of syntactic and syntactic-semantic behavior of verbs is a substantial task for linguists. Theoretical aspects of valency are challenging. Moreover, valency information stored in a lexicon (as valency properties are multifarious and cannot be described by general rules) belongs to the core information for any rule-based task of NLP (from lemmatization and morphological analysis through syntactic analysis to such complex tasks as e.g. machine translation).

There are tens of different theoretical approaches, tens of language resources and hundreds of publications related to the study of verbal valency in various natural languages. It goes far beyond the scope of this paper to give an exhaustive survey of all these enterprises – [1] gives a survey and a short characteristics of the most prominent projects.

The present paper is structured as follows: in Section 2 we summarize the basic properties of the lexicon VALLEX, in Section 3 we describe the human-annotated data where corpus occurrences of selected verbs are assigned to valency frames, in Section 4 we report the experiment with automatic frame assignment.

2 Valency Lexicon of Czech Verbs VALLEX

The VALency LEXicon of Czech verbs (VALLEX in the sequel) is a collection of linguistically annotated data and documentation, resulting from an attempt at formal

* The research reported in this paper has been partially supported by the grant of Grant Agency of Czech Republic No. 405/04/0243 and by the projects of Information Society No 1ET100300517 and 1ET101470416.

V. Matoušek et al. (Eds.): TSD 2005, LNAI 3658, pp. 99–106, 2005.

description of valency frames of Czech verbs. VALLEX version 1.0 was publicly released in autumn 2003[1]. VALLEX 1.0 contained roughly 1400 verbs with 4000 valency frames. At this moment, the latest version of VALLEX data contains roughly 2500 verbs with more than 6000 valency frames. All verb entries are created manually. Manual annotation and accent put on consistency of annotation are markedly time consuming and limit the speed of quantitative growth, but guarantees a significant rise of quality.

VALLEX is closely related to Prague Dependency Treebank (PDT)[2]. Both PDT and VALLEX are based on Functional Generative Description of Czech (FGD), being developed by Petr Sgall and his collaborators since the 1960s (see [3], valency theory within FGD esp. in [4]). Applying the principles of FGD to a huge amount of data means a great opportunity to verify and expand the theory, to refine the functional criteria set up. The modification of 'classical' FGD valency theory is used as the theoretical background in VALLEX 1.0 (see [5] for a detailed description of the framework).

On the topmost level, VALLEX[3] consists of **word entries** corresponding to complex units, verb lexemes (the VALLEX entries for the verbs *odpovídat* and *odpovídat se* is shown in Figure 1). The particular word entry is characterized by the **headword lemma**, i.e. the infinitive form of the respective verb (including the reflexive particle if it exists) and its **aspect** (perfective, imperfective or biaspectual). The tentative term **base lemma** denotes the infinitive of the verb, excluding the reflexive particle (i.e. the output of a morphological analysis).

Each word entry is composed of a non-empty sequence of **frame entries** relevant for the headword lemma. The frame entries (marked with subscripts in VALLEX) roughly correspond to individual senses of the headword lemma. The particular word entry is characterized by a **gloss** (i.e. verb or paraphrase roughly synonymous with the given frame/sense) and by **example(s)** (i.e. sentence fragment(s) containing the given verb used with the given valency frame). The core valency information is encoded in the **valency frame**.

Each valency frame consists of a set of **valency members / frame slots**, each corresponding to an (either required or specifically permitted) complementation of the given verb. The information on a particular valency member includes the following points:

– **'Functor'** expresses the type of relation between the verb and its complementation[4]. Complementations are divided into (i) inner participants / arguments (like Actor, Patient and Addressee for the verb *přinést*$_1$ [to bring], as in *někdo*.ACT *přinese něco*.PAT *někomu*.ADDR [sbd brings st to sbd] or Actor, Patient and Effect for the verb *jmenovat*$_3$ [to nominate], as in *někdo*.ACT *jmenuje někoho*.PAT *něčím*.EFF

[1] `http://ckl.mff.cuni.cz/zabokrtsky/vallex/1.0/`

[2] However, VALLEX is not to be confused with a larger valency lexicon PDT-VALLEX created during the annotation of PDT, see [2]. PDT-VALLEX contains more verbs (5500 verbs), but only frames occurring in PDT (over 9000 frames), whereas in the more complex VALLEX the verbs are analyzed in all their meanings. In addition, richer information is assigned to particular valency frames.

[3] Detailed description can be found in [6].

[4] The complete list of functors used in VALLEX together with English examples can be found in [6].

odpovídat (imperfective)

1 odpovídat$_1$ ∼ odvětit [answer; respond]
- frame: ACT_1^{obl} ADDR_3^{obl} $\mathrm{PAT}_{na+4,4}^{opt}$ $\mathrm{EFF}_{4,aby,ať,zda,že}^{obl}$ MANN^{typ}
- example: *odpovídal mu na jeho dotaz pravdu / že ...* [he responded to his question truthfully / that ...]
- asp.counterpart: odpovědět$_1$ pf.
- class: communication

2 odpovídat$_2$ ∼ reagovat [react]
- frame: ACT_1^{obl} $\mathrm{PAT}_{na+4}^{obl}$ MEANS_7^{typ}
- example: *pokožka odpovídala na včelí bodnutí zarudnutím* [the skin reacted to a bee sting by turning red]
- asp.counterpart: odpovědět$_2$ pf.

3 odpovídat$_3$ ∼ mít odpovědnost [be responsible]
- frame: $\mathrm{ACT}_1^{obl}\mathrm{ADDR}_3^{obl}\mathrm{PAT}_{za+4}^{opt}\mathrm{MEANS}_7^{typ}$
- example: *odpovídá za své děti; odpovídá za ztrátu svým majetkem* [she is responsible for her kids]

4 odpovídat$_4$ ∼ být ve shodě [match]
- frame: $\mathrm{ACT}_{1,že}^{obl}\mathrm{PAT}_3^{obl}\mathrm{REG}_7^{typ}$
- example: *řešení odpovídá svými vlastnostmi požadavkům* [the solution matches the requirements]

odpovídat se (imperfective)

1 odpovídat se$_1$ ∼ být zodpovědný [be responsible]
- frame: $\mathrm{ACT}_1^{obl}\mathrm{ADDR}_3^{obl}\mathrm{PAT}_{z+2}^{obl}$
- example: *odpovídá se ze ztrát* [he answers for the losses]

Fig. 1. VALLEX entries for the base lemma *odpovídat (answer, match)*.

[sbd nominates sbd as sbd]) and (ii) free modifications (adjuncts) as Time, Location, Manner and Cause[5].

- Possible **morphemic form(s)** – each complementation can be expressed by a limited set of morphemic means (pure or prepositional cases, subordinated clauses or infinitive constructions are the most important); possible morphemic form(s) are specified either explicitly (as a list of forms attached to a particular slot) or implicitly[6].
- **'Type'** – the following types of complementations are distinguished: obligatory (in the deep (tectogrammatical) structure) and optional for inner participants ('obl' and 'opt'), and obligatory and typical ('typ') for free modifications.

In addition to this obligatory information, also optional attributes may appear in each frame: flag for idiom, list of aspectual counterpart(s), information on control, affiliation to a syntactic-semantic class:

[5] Here we are leaving aside a small group of complementations on the border-line between inner participants and free modifications, quasi-valency complementations, see [5].

[6] The set of possible forms is implied by the functor of the complementation, see [6].

- **Flag for idiom** – VALLEX describes primary or usual meanings of verbs, however some very frequent idiomatic frames[7] are included as well. They are marked by idiomatic flag and include lemmas of words in the phraseme.
- **Aspectual counterpart** – aspectual counterpart(s) need not be the same for all senses of the given verb; if they exist, they are listed in particular frame entries[8] (see figure 1).
- **Control** – if a verb has a complementation in an infinitive form (regardless its functor), the valency member of the head verb that would be the subject of this infinitive is marked.
- **Syntactic-semantic classes** – particular frame entries are tentatively sorted into classes. Constructed in a 'bottom-up way', these classes are based on deep analysis of mainly syntactic properties of verbs in their particular senses. For the time being, 24 big groups involving next to half of the verb frames have been established[9].

3 VALEVAL

VALEVAL[10] is a lexical sampling experiment with VALLEX 1.0 for which 109 base lemmas from VALLEX 1.0 were selected. For each lemma 100 random sample sentences were extracted from CNC. See [7] for more details and examples.

Three human annotators in parallel were asked to choose the most appropriate verb entry and the frame for the extracted sentence within a context of the three preceding sentences. The annotators had also an option to indicate that the particular sentence is not a valid example (e.g. due to a tagging error) of the annotated lemma at all or that they got completely confused by the given context. A valid answer indicates a verb entry and a frame entry index. Optionally, a remark that the corresponding frame was missing could have been given instead of the frame entry index. If the annotators were not able to decide on a single answer, they have been given the possibility of assigning more than one valid answer (labelled as 'Ambiguous annotations' in Table 1). Also, a special flag could be assigned to a valid answer to indicate that the annotator is not quite sure (labelled as 'Uncertain annotations').

3.1 Inter-annotator Agreement

Table 2 summarizes inter-annotator agreement (IAA) and Cohen's κ statistic [9] on the 10256 annotated sentences. The symbol Ø indicates plain average calculated over base lemmas, wØ stands for average weighted by frequency observed in CNC. Considering all the three parallel annotations, the exact match of answers reaches 61% (weighted)

[7] Idiomatic frame is tentatively characterized either by a substantial shift in meaning (with respect to the primary sense), or by a small and strictly limited set of possible lexical values in one of its complementations.

[8] Iterative verbs occur in entries of the corresponding non-iterative verbs, but they have no own word entries.

[9] However rough these classes are, they serve for controlling the consistency of annotation.

[10] Inspired by SENSEVAL ([8]), a word sense disambiguation task, VALEVAL aims at valency frame disambiguation.

Table 1. Annotated data size and overall statistics about the annotations.

Lemmas annotated	109
Sentences annotated	10256
Parallel annotators	3
Total annotations	30765 (100%)
Uncertain annotations	1045 (3.4%)
Ambiguous annotations	703 (2.3%)
Marked as invalid example	172 (0.6%)
Annotator got confused	90 (0.3%)
Marked as missing frame	1673 (5.4%)

or 67% (unweighted). If the 'uncertainty' flags are disregarded, we find out that the agreement rises to 66% or 70%, respectively. In other words, annotators agree on the most plausible answer, even if they are not quite sure. If only such sentences where none of the annotators doubted are taken into account, the exact match reaches 68% or 74% (this comprises 90.5% of the sentences).

The κ statistic compensates IAA for agreement by chance. The level of 0.5 to 0.6 we achieve is generally considered as a *moderate agreement*, while 0.6 to 0.8 represents *significant agreement*. This moderate agreement is not an unsatisfactory result compared to other results such as [10], who reports pairwise IAA for French verbs between 60% and 65% and κ of 0.41.

Table 2. Inter-annotator agreement and κ.

	Match of 3 Annotators				**Average Pairwise Match**			
	IAA [%]		κ		IAA [%]		κ	
	wØ	Ø	wØ	Ø	wØ	Ø	wØ	Ø
Exact	61.4	66.8	0.52	0.54	70.8	74.8	0.54	0.54
Ignoring Uncertainty	65.9	69.8	0.58	0.59	74.8	77.7	0.60	0.59
Where All Were Sure	68.2	73.7	0.58	0.62	76.7	80.9	0.61	0.64

Average pairwise IAA is provided to allow for a rough comparison with some cited results, although the specific circumstances are not always directly comparable. [11] achieve an IAA for Czech verbs of 45% to 64%. For Japanese verbs, IAA of 86.3% is achieved by [12]. [13] report IAA of 71% for Senseval-2 English verbs tagged with WordNet synsets. Grouping some senses together to form a more coarse grained sense inventory allowed the authors to improve the IAA to 82%.

4 Automatic Frame Disambiguation

4.1 Data Source: 'Golden VALEVAL'

VALLEX frames correspond to verb senses (meanings). From this perspective, performing word sense disambiguation (WSD) of Czech verbs means choosing the most

Table 3. Baselines for WSD on 8066 'Golden VALEVAL' sentences for 108 lemmas.

	wØ	Ø
Entropy	1.54	1.28
VALLEX frames per lemma	12.46	7.61
Seen frames per lemma	5.85	4.85
10-fold Baseline WSD Accuracy	59.79%	66.19%

appropriate frame. 'Golden VALEVAL' is a corpus suitable for evaluating frame disambiguation. It comprises 8066 VALEVAL sentences covering 108 base lemmas where there was exact agreement across the annotators or a single answer was selected in a postprocess annotation aimed at eliminating clear typing errors and misinterpretations.

The difficulty of the WSD task is apparent from Table 3 looking at the (weighted or unweighted average) number of available frames per base lemma and entropy. The number of frames per lemma is estimated both from the whole VALLEX ('VALLEX frames per lemma') as well as from the set of actually observed frames in the golden VALEVAL corpus ('Seen frames per lemma').

The baseline accuracy is achieved by choosing the most frequent frame for a given lemma. The baseline was estimated by a 10-fold cross-validation (the most frequent frame is learned from 9/10 of the data and the unseen 1/10 is used to estimate the accuracy, the average result from 10 runs of the estimation is reported).

For purposes of further experiments, Golden VALEVAL was automatically tagged, lemmatized and enriched with surface syntactic structures automatically assigned by the Czech version of the parser reported in [14]. After the exclusion of unparsed sentences, 6666 sentences remained for our task.

4.2 Method and Selected Features

For an automatic selection of the VALLEX frame to which a given verb occurrence belongs, we generated a vector of features for each occurrence. We evaluated the decision tree machine learning method available in C5 toolkit[11]. 10-fold cross-validation was used for evaluation.

We experimented with several features containing information about the context of the verb. The following list describes different groups of features:

- Morphological: purely morphological information about lemmas in a 5-word window centered around the verb. Czech positional morphological tags (used also in PDT) contain 15 categories and all of these were taken as individual features, counting 75 features altogether.
- Syntax-based: information gained from the dependency tree of the sentence, including mostly Boolean information about morphological and lexical characteristics of dependent words (e.g. presence of a noun or a nominative pronoun in a given case dependent on the verb, presence of a given preposition with a given case dependent on the verb).

[11] `http://www.rulequest.com/see5-info.html`

4.3 Results

Weighting the accuracy by the number of sentences in our training set (labelled as Ø in Table 4), we gained 73.9% accuracy for morphological features and 78.5% accuracy for syntax-based features, respectively, compared to baseline 67.9% (baseline for the 6666 parsed sentences). Weighting the accuracy by the lemma frequency observed in the Czech National Corpus (labelled as wØ), the accuracy dropped to 67.1% for the morphological features and 70.8% for syntax-based features respectively, compared to baseline 63.3%.

Table 4. Accuracy of frame disambiguation.

	wØ	Ø
Baseline	63.3%	67.9%
Morphological	67.1%	73.9%
Syntax-based	70.8%	78.5%

The syntax-based features alone led to better results, and even the combination of both of the types of features did not bring any improvement. This could happen because the morphological information is already included in the syntax-based features (as they contain information mainly about morphological characteristics of syntactically related words) and because the syntactic structure of the sentence depicts enough information to achieve the rate of disambiguation which can be obtained using this method.

5 Conclusions and Future Work

We have presented the current state of building valency lexicon of Czech verbs VALLEX. We have also described the VALEVAL experiment which allowed us to improve consistency of selected VALLEX entries and provided us with golden standard data for WSD task. The first results in WSD are reported.

In future we plan to extend VALLEX in both qualitative aspects (e.g. description of alternations and types of reflexivity) and quantitative aspects. We will continue the WSD experiments, we intend to incorporate features based on WordNet classes and animacy.

References

1. Žabokrtský, Z.: Valency Lexicon of Czech Verbs. PhD thesis, Faculty of Mathematics and Physics, Charles University in Prague (2005) in prep.
2. Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V., Pajas, P.: PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In: Proceedings of The Second Workshop on Treebanks and Linguistic Theories. Volume 9 of Mathematical Modeling in Physics, Engineering and Cognitive Sciences., Vaxjo University Press (2003) 57–68

3. Sgall, P., Hajičová, E., Panevová, J.: The Meaning of the Sentence in Its Semantic and Pragmatic Aspects. D. Reidel Publishing Company, Dordrecht (1986)
4. Panevová, J.: Valency Frames and the Meaning of the Sentence. In Luelsdorff, P.L., ed.: The Prague School of Structural and Functional Linguistics, Amsterdam-Philadelphia, John Benjamins (1994) 223–243
5. Lopatková, M.: Valency in the Prague Dependency Treebank: Building the Valency Lexicon. Prague Bulletin of Mathematical Linguistics **79–80** (2003) 37–60
6. Žabokrtský, Z., Lopatková, M.: Valency Frames of Czech Verbs in VALLEX 1.0. In: Frontiers in Corpus Annotation. Proceedings of the Workshop of the HLT/NAACL Conference. (2004) 70–77
7. Bojar, O., Semecký, J., Benešová, V.: VALEVAL: Testing VALLEX Consistency and Experimenting with Word-Frame Disambiguation. Prague Bulletin of Mathematical Linguistics **83** (2005)
8. Edmonds, P.: Introduction to Senseval. ELRA Newsletter **7** (2002)
9. Carletta, J.: Assessing agreement on classification task: The kappa statistics. Computational Linguistics **22** (1996) 249–254
10. Véronis, J.: A study of polysemy judgements and inter-annotator agreement. In: Programme and advanced papers of the Senseval workshop, Herstmonceux Castle (England) (1998) 2–4
11. Hajič, J., Holub, M., Hučínová, M., Pavlík, M., Pecina, P., Straňák, P., Šidák, P.: Validating and Improving the Czech WordNet via Lexico-Semantic Annotation of the Prague Dependency Treebank. In: Proceedings of LREC 2004. (2004)
12. Shirai, K.: Construction of a Word Sense Tagged Corpus for SENSEVAL-2 Japanese Dictionary Task. In: Proceedings of LREC 2002. (2002) 605–608
13. Babko-Malaya, O., Palmer, M., Xue, N., Joshi, A., Kulick, S.: Proposition Bank II: Delving Deeper. In: Frontiers in Corpus Annotation. Proceedings of the Workshop of the HLT/NAACL Conference. (2004) 17–23
14. Charniak, E.: A Maximum-Entropy-Inspired Parser. In: Proceedings of NAACL-2000, Seattle, Washington, USA (2000) 132–139

Chapter B

Theoretical Aspects of Valency and Their Manifestation in the Lexicon

B.1 Recent Developments in the Theory of Valency in the Light of the Prague Dependency Treebank

In Šimková, M. (ed.) *Insight into Slovak and Czech Corpus Linguistics*, p. 83-92. Bratislava: Veda, 2006
(with co-author J. Panevová)

Recent Developments in the Theory of Valency in the Light of the Prague Dependency Treebank*

MARKÉTA LOPATKOVÁ, JARMILA PANEVOVÁ

1 THE FRAMEWORK

The Functional Generative Description (FGD, see Sgall, 1967, Sgall et al., 1986) was applied as a general framework for the development of the valency theory (see Panevová, 1974-75, 1980, 1994) as well as for the design of the Czech syntactically annotated corpus (PDT, see Hajič, 1998, Hajičová et al., 2001).

Valency is understood as a lexico-syntactic attribute of a word – more precisely, of a particular lexical sense of the lemma, called here lexis ("lexie" in Czech terminology, see Filipec and Čermák, 1985). More precisely, we can understand a lexis as a pair formed by a lexical unit and one of its meanings.[1] A valency frame (VF) is assigned to every auto-semantic lexical unit (lexis). This, however, may be empty, e.g. with the Czech verb *pršet [to rain]*, with nouns such as *stůl [the table],* adjectives as *hezký [beautiful]*. The labels used for the valency slots belong to the underlying structure (tectogrammatics) and, together with the lexical unit (lexis), they constitute a tectogrammatical representation of the lexical entry. With regard to the applied tasks, we include the morphemic counterparts of the particular valency slots as a part of the (complex) frame of the given unit.

Valency is prototypically connected with verbs. We have distinguished two main classes of verbal complements:

(i) inner participants, IP in the sequel (ACT(or), PAT(ient), ADDR(essee), ORIG(in) and EFF(ect)),
(ii) free modifications, FM in the sequel.

The criteria for the distinction between these two classes are given in Panevová (quoted above).

Valency frames of lexes are constituted by their respective inner participants (either obligatory or optional) and by their obligatory free modifications.[2]

* The work reported on in this paper has been carried out under the project of "Centers of Excellence" supported by MŠMT, grant No LN00A063. It has been partly supported from the grant GAČR No 405/04/0243.

[1] The formal representation of lexis in FGD has not yet been specified. The surface shape (lemma) of the lexical item is used instead (with a differentiating subscript, if necessary).

[2] We prefer this terminology rather than the terminology used in Daneš et al., 1981 and "Mluvnice češtiny 3", 1987. There the term "potenciální" (potential) is used for optional as well as for obligatory positions of VF omitted on the surface. Moreover, the difference between the VF as a part of lexicon and its application for the concrete utterance is not reflected in the terminology common in Czech handbooks.

We share Tesnière's (1959) approach as to the one-argument and two-argument verbs: the first slot is structured as ACT(or) (though it corresponds to different semantic (ontological) roles, such as Bearer, Processor, Stimulus etc.); with two-argument verbs the inner participants are structured as ACT(or) and PAT(ient). The relation between the syntactic arguments and their cognitive roles is called a "shifting of participants", see Panevová, 1980. If the verb has three (or more) valency slots, the semantics of them is taken into account. This strategy agrees with the theory of case meanings, distinguishing between syntactic (grammatical) cases and semantic (concrete) cases (see Kuryłowicz, 1949): the valency slots of ACT and PAT are occupied mostly by syntactic cases (Nominative and Accusative, respectively), while the other participants and free modifications are expressed mostly by cases with concrete (semantic) meanings.

2 AN INTRODUCTION OF QUASI-VALENCY COMPLEMENTS

In section 1 we briefly summarized the basic features of our valency theory of verbs. However, in the course of empirical studies of material, especially in connection with the building of the valency lexicon of verbs VALLEX (see Lopatková, Žabokrtský, 2003 and section 5 below) and with a tectogrammatical annotation of PDT (see Urešová, this volume), some unresolved problems appeared. Firstly, it was necessary to introduce some additional functors (types of syntactic-semantic relations) for newly discovered semantically relevant distinctions (namely OBST(acle) and MED(iator)). In analyzing their semantic and syntactic distribution, we observed that they share partly the features of inner participants, and partly the features of free modifications. Secondly, revisiting the list of verbal complements introduced earlier, we discovered that some complements (namely DIFF(erence) and INT(ent)) also share important features of inner participants (see (i), (ii) and (iii)), although they also have some of the characteristic features of free modifications (see (iv), (v) and (vi)):

(i) they are governed (their morphemic shape is determined) by their verbal heads
(ii) they occur with a limited class of verbs
(iii) they cannot be repeated,

however

(iv) as to their meaning, they are semantically homogeneous
(v) they do not underlie the "shifting"
(vi) they are mostly optional.

We also reconsidered the complements ADDR, ORIG (and perhaps EFF) from this point of view. The complements ADDR and ORIG undoubtedly fulfill (i), (ii), (iii) characteristics for IP, but also (iv),[3] which is typical of FM; they do not meet (v) and (vi). The features of EFF shared with quasi-valency complements are limited; (i), (ii) and (iii) are present in EFF, but one of the most important quasi-valency features (iv) is missing here. This is the main reason why we still classify EFF as an inner participant. However, we are still undecided as to whether the ADDR and ORIG should not be classified as quasi-valency complements, too.

2.1 OBSTACLE

The meaning of **OBST(acle)** is expressed in Czech by the prepositional group *o + Accusative* with verbs like *zakopnout [to stumble], uhodit se [to strike oneself], bouchnout se [to bump oneself], zranit se [to injure oneself], píchnout se [to prick oneself], bodnout se [to prick oneself]*.

[3] This statement is valid at least for verbal valency features. As for nouns, see Section 4 below.

Their form is governed by their head verbs. In handbooks on Czech syntax they are classified as Means (Instrument), but they undoubtedly have a special instrumental semantics, see (1), (2) and (3):

(1) Jan zakopl nohou o stůl
[John stumbled over the table with his leg]
(2) Matka se píchla nůžkami
[Mother pricked herself with the scissors]
(3) Růženka se píchla o trn
[Sleeping Beauty pricked herself on a thorn]

In (1) *noha [leg]* is a proper means (Instrument), while the construction *o stůl [about the table]* is not. In (2) *nůžky [scissors]* refers to a device used as an Instrument proper, its semantics includes the semantics of movement with this instrument. In (2) the manipulation with scissors is presumed, while in (3) the noun *trn [thorn]* (with an instrumental semantics) is fixed (see also Apresjan, 2001). The feature of an unconscious action is typical of (3), while in (2) the action can be either conscious or unconscious. For the semantics of "fixed" Instrument (expressed by the prepositional group *o + Accusative*) the new label Obstacle was proposed (initially in Panevová, 2003). All the verbs listed in this sample imply their unconsciousness. The verbal modification of **Obstacle** shares the features of the group of inner participants (i), (ii) and (iii), but also all the features listed above as free modification attributes (iv), (v), and (vi)[4].

2.2 MEDIATOR

Also, the Czech prepositional group *za + Accusative* is described in syntactic handbooks as a kind of Instrument, see e.g. (4), (5), (6):

(4) Otec přitáhl kluka levou rukou za ucho
[Father has drawn boy's ear by his left hand]
(5) Když jsem odcházel, zatahal mě soused za rukáv
[When I was leaving, the neighbor pulled my sleeve]
(6) Jan přivedl psa za obojek
[John brought the dog by its collar]

Examples (4) to (6) demonstrate that the semantics of this prepositional group is different from the pure Instrument. Pure Instrument is usually used by the Actor of the action directly, while in (4) to (6) the instrument is a part of another entity (the ear belongs to the boy in (4) and as a part of a boy it is used for drawing the boy). In (4) the Instrument proper is present (*ruka [hand]*). The Actor uses his own hand as a means to reach the boy, and he uses the boy's ear as a **Mediator** for reaching him. Like the Obstacle, the Mediator shares some features of IP and some of the class of FM. Unlike the Obstacle, we have not yet found any verb with an obligatory Mediator.

2.3 DIFFERENCE

The prepositional group *o + Accusative,* although it mostly combines with the comparatives of adjectives or adverbs, can also occur with some verbs (see e.g. (7), (8), (9) for verbs, (10) for an adverb):

[4] Feature (vi) has some exceptions: we have found the verbs *zavadit [to touch], (za)chytit (o něco) [to get caught (on st)]* with obligatory OBST.

(7) Inflace se zvýšila proti roku 2000 o několik procent.
[The inflation has increased in comparison with 2000 by several percent]
(8) Náš tým zvítězil o dvě branky
[Our team won by two goals]
(9) Jan zvítězil v závodě o prsa
[John won the race by a hair's breadth]
(10) Postupte o dva schody výš
[Move two steps higher]

The modification of **DIFF(erence)** can be characterized as a kind of extent, but while the general extent expresses nothing more than a high or low degree, the modification of DIFF specifies the extent more precisely. At least two entities are compared here, although one of them is more or less implicit (inflation in the current year and in 2000 are compared in (7), the score of a match of two teams are compared in (8), John's rivals are understood in (9) as the other entity) and the difference between them is explicitly expressed by the Difference modification.

2.4 INTENT

The modification of **INT(ent)** is compatible mainly with the verbs of motion and it differs from the FM of AIM: an actor of the INT is identical with the person that provides the intended action himself/herself (the action can be transformed into a nominalization, see e.g. (12), contrary to (13), where the FM of AIM is expressed). The actor (mother in the case of (13)) only transfers potatoes from one place to another. The difference between INT and AIM could be exemplified by the acceptability of (14a) and unacceptability of (14b).[5]

(11) Jan se šel koupat
[John went to swim]
(12) Helena šla na jahody
[Helen went (to pick) strawberries / *lit.* Helen went on strawberries]
(13) Matka šla do sklepa pro brambory
[Mother went to the cellar for potatoes]
(14a) Helena šla do krámu pro jahody
[Helen went to the shop for strawberries]
(14b) *Helena šla do krámu na jahody
[*Helen went to the shop (to pick up) strawberries / *lit.* Helen went to the shop on strawberries]

3 VALENCY OF ADJECTIVES

Our analysis of adjective valency was aimed at the verification of two hypotheses:

(i) that the valency slots of adjectives share the roles of verbal complements;
(ii) that the shifting of participants is here valid in the same manner as with verbs (with one natural exception: one of the valency slots is absorbed by the governing noun in

[5] The introduction of the INT complement is supported by the findings presented in Poldauf, 1959. The prototypical expression of an INT is an infinitive; unprototypically, the prepositional expression is used (see (12)); it implies the active participation of the actor in collecting strawberries. This is the reason why (14b) is meaningless (at least in our actual world), somebody else (other than Helen) has collected the strawberries and delivered them to the shop.

noun phrases or by the subject position in the clauses with the copula *být [to be]* so it is excluded from the valency frame of the respective adjective).

In the case of primary adjectives, the position of ACT is absorbed; with deverbal adjectives the absorbed position depends on the type of derivation (with active participles the position of ACT is absorbed as well, with passive participles PAT, ADDR or EFF is absorbed, for details see Panevová, 1998).

Otherwise, the deverbal adjectives share the valency of their source verbs.

The question of the lexical ambiguity of adjectives used for human qualities remains open. This consideration concerns such adjectives as *hrdý [proud], věrný [faithful]* etc. They are used either as the "absolute" attribute of a noun (and they have an empty valency frame), or they are used as relative adjectives with an obligatory PAT *(hrdý na* + Acc*, věrný* + Dat). We have also considered an alternative solution, where we have to deal with a single lexical sense for absolute and relative usage and where the optional PAT enters their valency frame (for more examples, see Panevová, 1998 and Panevová, in prep.).

4 VALENCY OF NOUNS

The set of valency complements of nouns was extended, as proposed by Piťha, 1981, if compared with the set of valency complements of verbs. We have accepted his proposal as to the complements called there **MAT(erial)** (as an obligatory or an optional noun participant) and **APP(urtanance**) (as a free noun modification, obligatory with the listed nouns). We have reconsidered his proposal to classify **ID(entity)** as an optional participant of a noun; it should belong to the class of FM, because any noun can have its name (not only *loď Titanic [boat Titanic],* but also *tužka Koh-i-nor [pencil Koh-i-nor], souprava Julie [set Julia])*.

In the valency frame of many nouns, the same complements occur as in the VF of verbs. This is obvious for deverbal nouns (for details see Novotný, 1980, Karlík, 2000, Panevová, 2000 and esp. Řezníčková-Kolářová, 2003, Kolářová, in prep.). Moreover, the complements (functors) typical of verbs are compatible with a high number of primary nouns (e.g. PAT in *názor na něco [opinion on], příklad na něco/něčeho [example for], kniha o něčem [book on],* ADDR in *dárek někomu [gift to]*, ORIG in *daň z pozemku [tax for]*). In the last two cases, we again perhaps have to do with the absorption of one participant built within the head noun (*dárek* and *daň* are patients themselves, a gift is what was given, tax is what is paid).

The functor called ORIG(in) has a special position among noun complements. Although it has its counterpart within verbal inner participants, with nouns it typically behaves as a free modification: it is compatible with any primary noun and it can be repeated (*šaty ze lnu od starší sestry [a dress from linen from my elder sister], nábytek ze dřeva od našeho hlavního dodavatele [furniture from wood from our main provider]).* The interpretation of the inanimate noun expressing an Origin is material, while an animate name (and its equivalents as the names of institutions, human collectives etc.) corresponds to the source. A re-classification of Origin as a FM noun complement – proposed here for the first time within our framework – is based on its syntactic behaviour with nouns (different from its behaviour with verbs, where it cannot be repeated and it is not compatible with every verb).

5 THE BUILDING OF A VALENCY LEXICON BASED ON THE THEORY DESCRIBED

A description of valency is impossible without a good syntactically based framework, and – since valency differs from one lexical item to another – it cannot be described by general rules. Therefore a valency lexicon belongs among the basic language resources indispensable

for any rules-based task of NLP (Natural Language Processing). Here we refer to the valency lexicon VALLEX, which has been created in connection with the annotation of PDT.[6]

The Valency Lexicon of Czech Verbs, Version 1.0 (VALLEX 1.0, http://ckl.mff.cuni.cz/zabokrtsky/vallex/1.0/) is a collection of linguistically annotated data and documentation, resulting from the attempt at formal description of the valency frames of Czech verbs. VALLEX 1.0 contains roughly 1400 verbs in all their senses (app. 4000 frame entries / senses). VALLEX is designed both for human readers and for application tasks in NLP as e.g. machine translation or information retrieval.

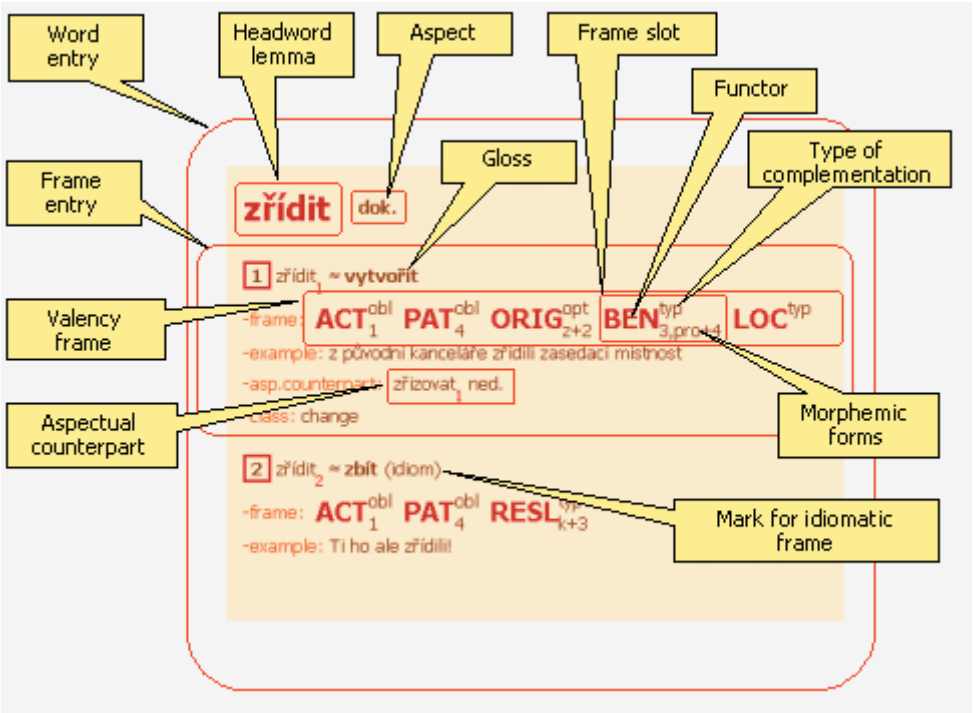


Figure 1: Word entry in VALLEX

A Czech verb as a whole, a verb lexeme (**word entry** in VALLEX) is an abstract unit made up by all the senses of a particular verb. A word entry consists of a (non-empty) sequence of **frame entries**, each of which corresponds to a single sense ("lexis", see above). Each frame entry describes the valency frame itself, the specification of a sense in question (by gloss(es) and example(s)), and additional information (as e.g. aspect, type of reflexivity, control, (preliminary) semantic class). A **valency frame** itself is a sequence of **frame slots** corresponding to (either required or specifically permitted) complements of a given verb. Each valency slot is characterized by its **functor**, i.e. the name of the syntactic-semantic

[6] Besides VALLEX, a larger valency lexicon (called PDT-VALLEX, see e.g. Hajič et al., 2003, Urešová, this volume) has been created during the annotation of PDT. PDT-VALLEX contains more verbs (5200 verbs), but with only those of their senses that occurred in PDT, whereas in VALLEX the verbs are analyzed in their full complexity, in all their senses. In addition, richer information is assigned to particular valency frames in VALLEX, and stress is laid on the consistency and completeness of annotation.

relation (labels of underlying roles), and the possible morphemic form(s) (specification of morphemic case, prepositional group, infinitive or subordinated verbal construction).

A word entry in VALLEX corresponds to the whole lexeme; it consists of a (non-empty) sequence of frame entries corresponding to a single sense.

We have formulated the following principles and functional criteria for distinguishing particular senses adopted that are connected with their valency. The principles can be characterized by two statements:

A. any change in valency frame (either in functor, in the combination of functors, or possible form(s) of functor) justifies an introduction of a new frame entry;

B. any significant change in sense justifies the introduction of a new frame entry.

These fundamental principles imply the following rules.

(i) The difference in the sense is a necessary but not sufficient condition for a postulation of two (or more) valency frames – a (slight) difference in the sense is ignored if lexical units do not differ syntactically.

(15) *hýbat*$_1$ *[to move]*[7] … ACT(1;obl) PAT(Instr,s+Instr;obl)
hýbat rukou; hýbat (s) křeslem
[to move (with) sb`s hand, to move an armchair]

In Czech lexicons "Slovník spisovného jazyka českého" [The dictionary of Standard Czech] (1964) as well as in "Slovesa pro praxi" [Verbs for Practice] (1997) two distinct senses are distinguished – "uvádět něco v pohyb, pohybovat" [to set st in movement, to move st] and "měnit polohu" [to change position (of st)]. In VALLEX, these two usages of the verb *hýbat* in (15) are described in a single valency frame – the difference in the senses is not taken into account, their syntactic behaviour being the same. The decision to ignore this type of difference is based on the fact that such a "fine-grained" distinction of senses is not reflected in the syntactic behaviour of the given lexical units and they are often not perceived, even by a human reader in real texts.

(ii) Two different senses can have an identical valency frame.

(16a) *chovat*$_1$ *[to cradle]* … ACT (1;obl) PAT(4;obl)
chovat dítě (v náručí)
[to cradle a child (in one`s arms)]

(16b) *chovat*$_2$ *[to keep]* … ACT (1;obl) PAT(4;obl)
chovat prasata (na farmě)
[to keep pigs (on a farm)]

The indisputable different senses of the verb *chovat* have the same valency frame consisting of two inner participants, Actor and Patient with the same morphemic forms; however, the difference of the sense has to be reflected by distinguishing two different frame entries in VALLEX.

(iii) The change in morphemic realization signalizes the possibility of different senses.

(17a) *hlásit se*$_2$ *[to be counted among sb*] … ACT(1;obl) PAT(k+3;obl)
hlásit se ke komunistům

[7] The lower numeral index attached to the lemma denotes a particular frame entry in VALLEX notation.

[to be counted among communists]

(17b) *hlásit se*$_4$ *[to apply for st]* … ACT(1;obl) PAT(o+4;obl)
hlásit se o svá práva
[to apply for own rights]

The change in morphemic realization signalizes different senses and thus two lexical items *hlásit se*$_2$ and *hlásit se*$_4$ are distinguished.

(iv) On the other hand, a particular complement in a valency frame can have morphemic variants (if they differ stylistically, rather than in their semantics).

(18) *učit*$_1$ *[to teach]* … ACT(1;obl) ADDR(4;obl) PAT(3,4,inf,že,zda,aby,jak;obl)
Učitel učí žáky matematice / matematiku / pracovat / ...
[Teacher teaches his pupils mathematics$_{Dat}$ / mathematics$_{Acc}$ / to work / ...]

With this lexical unit there is more than a single possibility to express the obligatory Patient.

(v) A change in valency frame is connected with a change of sense – two valency frames cannot share their senses.

(19a) *postavit*$_1$ *[to raise]* … ACT(1;obl) PAT(4;obl)
postavit sloup
[to raise a column]

(19b) *postavit*$_2$ *[to build]* … ACT(1;obl) PAT(4;obl) ORIG(z+2;opt)
postavit budovu; postavit model letadla z balzy
[to build up a building; to construct a model of a plane from balsa wood]

(20a) *poslat*$_1$ *[to send]* … ACT(1;obl) ADDR(3;obl) PAT(4;obl)
poslat matce dárek k narozeninám.
[to send sb`s mother a birthday gift]

(20b) *poslat*$_2$ *[to send]* … ACT(1;obl) PAT(4;obl) DIR3(;obl)
poslat zásilku do Konga
[to send a consignment to Congo]

The valency frames in (19a) and (19b) differ in the presence of an optional inner participant ORIG(in) – *postavit*$_1$ *[to raise]* cannot be modified by this complement. This distinction entails a clear distinction in the senses of *postavit*$_1$ and *postavit*$_2$ (reflected also by different translation equivalents, *to raise* and *to build*).

With some groups of verbs this principle is not obvious at first sight – they have two valency frames and their sense is rather close, e.g. *poslat* in (20a) and (20b). However, the detailed analysis of syntactic and semantic properties of some of these groups given in Benešová, 2004 shows clear syntactic and semantic distinctions in sense between them.

(vi) Different valency frames can reflect a primary and a secondary (figurative) usage of a given verb.

(20a) *dopadnout*$_1$ *[to fall (down)]* … ACT(1;obl) DIR3(;obl)
dopadnout na zem
[to fall down to the ground]

(20b) *dopadnout*$_2$ *[to strike]* … ACT(1;obl) PAT(na+4;obl)
Dopadly na ně starosti.
[Troubles have fallen on them]

Directionality proper and directionality in a metaphorical sense are met in (20a) and (20b). Despite the same morphemic realizations, different functors, namely DIR3 (direction – to where) and PAT, are assigned to the second complement. This distinction is justified by different syntactic-semantic features (*dopadnout*$_1$ belongs to the "verbs of motion", unlike *dopadnout*$_2$).

Distinguishing the particular senses of a single verb lexeme is amongst the most complicated problems in the domain of constructing a lexicon. We have tried to discuss and exemplify the criteria connected with the valency behaviour of verbs.

6 CONCLUSION

The Czech data analyzed during the development of the PDT present some new issues not yet solved within the theoretical background. In confronting these issues, we have made some modifications in the framework: we have introduced new types of functors (syntactic-semantic relations) and we have shifted some functors into another class of valency complements. We have presented here several examples illustrating the methodology used in building up the valency lexicon (VALLEX 1.0). The relations between the lexical meanings of verbal units and their valency frames are illustrated in Section 5. We can conclude, however, that the changes to the framework resulting from the annotation of relatively large data are not substantial, although they have brought some refinements of the theory of FGD.

REFERENCES

Apresjan, J. D. (2001): Znachenije i upotreblenije. Voprosy jazykoznanija 4, pp. 3-22.

Benešová, V. (2004): Delimitace lexií českých sloves z hlediska jejich syntaktických vlastností. Diplomová práce, FFUK, Praha.

Daneš, F., Hlavsa, Z. a kol. (1981): Větné vzorce v češtině. Academia, Praha.

Filipec, J., Čermák, F. (1985): Česká lexikologie. Academia, Praha.

Hajič, J. (1998): Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In: Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová (ed. E. Hajičová), Karolinum, Charles University Press, Prague, pp. 106-132.

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 Second Workshop on Treebanks and Linguistic Theories, pp. 57-68. Vaxjo University Press.

Hajičová, E., Panevová, J., Sgall, P. (2001): Manuál pro tektogramatické značkování. Verze IV. Technická zpráva, ÚFAL MFF UK.

Karlík, P. (2000): Valence substantiv v modifikované valenční teorii. In: Čeština – univerzália a specifika 2, Sborník z konference ve Šlapanicích u Brna, Masarykova Univerzita, Brno, pp. 181-192.

Kolářová, V. (in prep.): Valence deverbativních substantiv v češtině. (Manuscript of PhD thesis).

Kuryłowicz, J. (1949): Le problème du classement des cas. Biulletyn Polskiego Towarzyszstva Językoznawczego, Vol. 9, pp. 20-43.

Lopatková, M., Žabokrtský, Z., Skwarska, K., Benešová, V. (2003): Valency Lexicon of Czech Verbs VALLEX 1.0. CKL/UFAL Technical Report TR-2003-18, 2003.

Mluvnice češtiny 3, Skladba (1987). Akademie, Praha.

Novotný, J. (1980): Valence dějových substantiv v češtině. In: Sborník Pedagogické fakulty v Ústí n. Labem, Praha.

Panevová, J. (1974-75): On Verbal Frames in Functional Generative Description. Part I, The Prague Bulletin of Mathematical Linguistics 22, pp 3-40, Part II, The Prague Bulletin of Mathematical Linguistics 23, pp. 17-52.

Panevová, J. (1980): Formy a funkce ve stavbě české věty. Academia, Praha.

Panevová, J. (1994): Valency Frames and the Meaning of the Sentence. In: The Prague School of Structural and Functional Linguistics (ed. Ph. L. Luelsdorff), Amsterdam-Philadelphia, John Benjamins, pp. 223-243.
Panevová, J. (1998): Ještě k teorii valence. In: Slovo a slovesnost 59, pp.1-14.
Panevová, J. (2000): Poznámky k valenci podstatných jmen. In: Čeština – univerzália a specifika 2, Sborník z konference ve Šlapanicích u Brna, Masarykova Univerzita, Brno, pp. 173-180.
Panevová, J. (2003): Some Issues of Syntax and Semantics of Verbal Modifications. In: Proceedings MTT 2003, First International Conference on Meaning-Text Theory, pp. 139-146. Ecole Normale Supérieure.
Panevová, J. (in prep.): Valence vybraných českých adjektiv ve světle ČNK. Slavistična revija.
Piťha, P. (1981): On the Case Frames of Nouns. Prague Studies in Mathematical Linguistics 7, Academia, Prague, pp. 215-224.
Poldauf, I. (1959): Děj v infinitivu. In: Slovo a slovesnost 20.
Řezníčková-Kolářová, V. (2003): Czech Deverbal Nouns: Issues of Their Valency in Linear and Dependency Corpora. In: Proceedings of the Workshop on Shallow Processing of Large Corpora (SProLaC 2003), pp. 88-97. UCREL, Lancaster University.
Sgall, P. (1967): Generativní popis jazyka a česká deklinace. Academia, Praha.
Sgall, P., Hajičová, E., Panevová, J. (1986): The Meaning of the Sentence in Its Semantic and Pragmatic Aspects (ed. by J. Mey), Dordrecht: Reidel and Prague: Academia.
Slovník spisovného jazyka českého (1964). Praha.
Slovesa pro praxi (1997): Svozilová, N., Prouzová, H., Jirsová, A. (autoři), Academia, Praha.
Urešová, Z. (this volume): The verbal valency in the Prague Dependency Treebank.
Tesnière, L. (1959): Eléments de syntaxe structurale. Paříž.

B.2 Valency Lexicon of Czech Verbs: Alternation-Based Model

In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, Vol. 3, p. 1728-1733, Paris, 2006. ELRA
(with co-authors Z. Žabokrtský and K. Skwarska)

Valency Lexicon of Czech Verbs: Alternation-Based Model

Markéta Lopatková, Zdeněk Žabokrtský, Karolina Skwarska

Institute of Formal and Applied Linguistics, Charles University, Prague
Malostranské náměstí 25, Prague 1, 118 00, Czech Republic
{lopatkova,zabokrtsky}@ufal.mff.cuni.cz

Abstract

The main objective of this paper is to introduce an alternation-based model of valency lexicon of Czech verbs VALLEX. Alternations describe regular changes in valency structure of verbs – they are seen as transformations taking one lexical unit and return a modified lexical unit as a result. We characterize and exemplify 'syntactically-based' and 'semantically-based' alternations and their effects on verb argument structure. The alternation-based model allows to distinguish a minimal form of lexicon, which provides compact characterization of valency structure of Czech verbs, and an expanded form of lexicon useful for some applications.

Introduction

The verb is traditionally considered to be the center of the sentence, and the description of syntactic and syntactic-semantic behavior of verbs is a substantial task for linguists. Theoretical aspects of valency are challenging. Moreover, valency information stored in a lexicon (as valency properties are diverse and cannot be described by general rules) belongs to the core information for any rule-based task of NLP (from lemmatization and morphological analysis through syntactic analysis to such complex tasks as e.g. machine translation).

There are tens of different theoretical approaches, tens of language resources and hundreds of publications related to the study of verbal valency in various natural languages. It goes far beyond the scope of this paper to give an exhaustive survey of all these efforts – Žabokrtský (2005) gives a survey and short characteristics of the most prominent projects (i.e. (Fillmore, 2002), (Babko-Malaya et al., 2004), (Erk et al., 2003) and (Mel'čuk and Zholkovsky, 1984)).

The present paper is structured as follows: in the first section the valency lexicon VALLEX is introduced. Section 2. deals with the concept of alternations – we present alternations as transformations that describe regular changes in the valency structure of verbs (and reduce lexicon redundancy). We characterize basic rules for their representation and exemplify basic types of alternations. Section 3. gives a brief sketch of minimal and expanded form of the lexicon.

1. Valency lexicon VALLEX

The valency lexicon VALLEX is a collection of linguistically annotated data and documentation, resulting from an attempt at a formal description of valency frames of roughly 4300 most frequent Czech verbs. It is closely related to Prague Dependency Treebank (PDT), see (Hajič, 2005).[1] VALLEX provides information on the valency structure of verbs in their particular meanings / senses, possible morphological forms of their complementations and additional syntactic information, accompanied with glosses and examples (briefly described below; the theoretical background of Functional Generative Description of Czech is presented in (Sgall et al., 1986) and (Panevová, 1994), its application on VALLEX is specified in (Lopatková, 2003)). All verb entries in VALLEX are created manually; manual annotation and accent put on consistency of annotation are highly time consuming and limit the speed of quantitative growth, but allow for reaching desired quality.

VALLEX version 1.0 was publicly released in autumn 2003. The second version of the lexicon, VALLEX 2.0, which adopted the alternation-based model will be available this autumn (2006) at http://ufal.mff.cuni.cz/vallex/.

1.1. Structure of VALLEX

VALLEX can be seen as having two components, a data component and a grammar component.

Formally, the **data component** consists of word entries corresponding to verb lexemes. Lexeme is an abstract twofold data structure which associates lexical form(s) and lexical unit(s) (see Fig. 1).

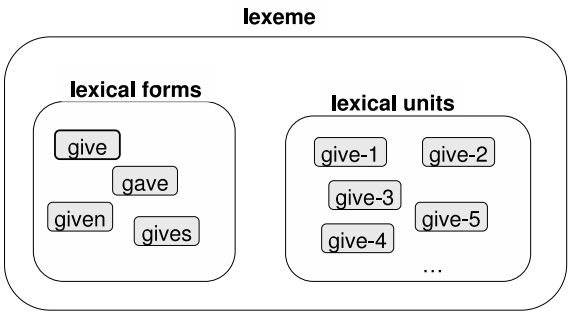


Figure 1: Lexeme, lexical form, and lexical unit.

Lexical forms are all possible manifestations of a lexeme in an utterance, as e.g. perfective, imperfective and iterative verb lemmas, all their morphological manifestations, reflexive and irreflexive forms etc. In the lexicon, all lexical

[1] However, VALLEX is not to be confused with a bit larger valency lexicon PDT-VALLEX created during the annotation of PDT, see (Hajič et al., 2003). PDT-VALLEX has originated as a set of valency frames instantiated in PDT, whereas in the more complex and more elaborated VALLEX verbs are analyzed in all their complexity.

forms of a lexeme are represented by perfective, imperfective and iterative infinitive forms (if they exist), the so called **(headword) lemma(s)**.

Concerning **lexical units (LUs)**, the concept introduced in (Cruse, 1986) has been adopted: LUs are "form-meaning complexes with (relatively) stable and discrete semantic properties". Particular lexical unit is specified by particular meaning / sense, loosely speaking,'given word in the given sense'.[2] Each lexical unit is characterized by a **gloss** (i.e. a verb or a paraphrase roughly synonymous with the given meaning / sense) and by **example(s)** (i.e. sentence fragment(s) containing the given verb used with the given valency frame). The core valency information is encoded in the **valency frame** consisting of a set of **valency members / slots**. Each of these valency members corresponds to an individual – either required or specifically permitted – complementation of the given verb (assigned with its possible morphological forms and a flag for obligatorness). In addition to this obligatory information, also optional attributes may appear in each LU: a flag for idiom, information on control, affiliation to a syntactic-semantic class and a list of alternations that can be applied to this LU (accompanied by examples as illustrated below), see Fig. 2.

The **grammar component** consists of a set of transformations that can be applied to particular LUs (as specified in the data component) to obtain derived LUs and thus an expanded form of the lexicon. These transformations explicitly cover possible alternation constructions for individual verb forms (they are described in more details in Section 2.2.).

1.2. Basic quantitative characteristics of VALLEX

VALLEX 2.0 contains almost 2100 lexemes. Valency frames of around 6350 LUs are stored in the lexicon. From the other point of view, it describes roughly 4300 verbs (counting perfective forms (ca 1950 verbs), imperfective forms (2250 verbs) as well as biaspectual forms (96 verbs); in addition to these numbers, VALLEX contain also 335 iterative verbs).

2. Alternations

When studying the valency of Czech verbs, it proves to be fruitful to exploit the concept of Levin's alternations (Levin, 1993) and to adapt it for Czech. Levin's alternations describe different changes in argument structure of lexical units. Though our main goal is rather different from that of Levin (Levin builds semantically coherent classes from verbs which undergo particular sets of alternations), the concept of alternations enables us to systematically describe regular changes in argument structure of verbs. Levin recognizes around 45 alternations for English (some of them with more variants). Similar behavior of verbs can be detected in Czech in spite of the typological character of this inflective language. Several of these alternations are described in Czech linguistic works, e.g. in (Daneš, 1985), (Mlu, 1987), (Panevová, 1999), but no Czech lexicon has reflected this model yet.

[2]This concept of LU corresponds to the Filipec's 'monosemic lexeme' as specified in (Filipec, 1994).

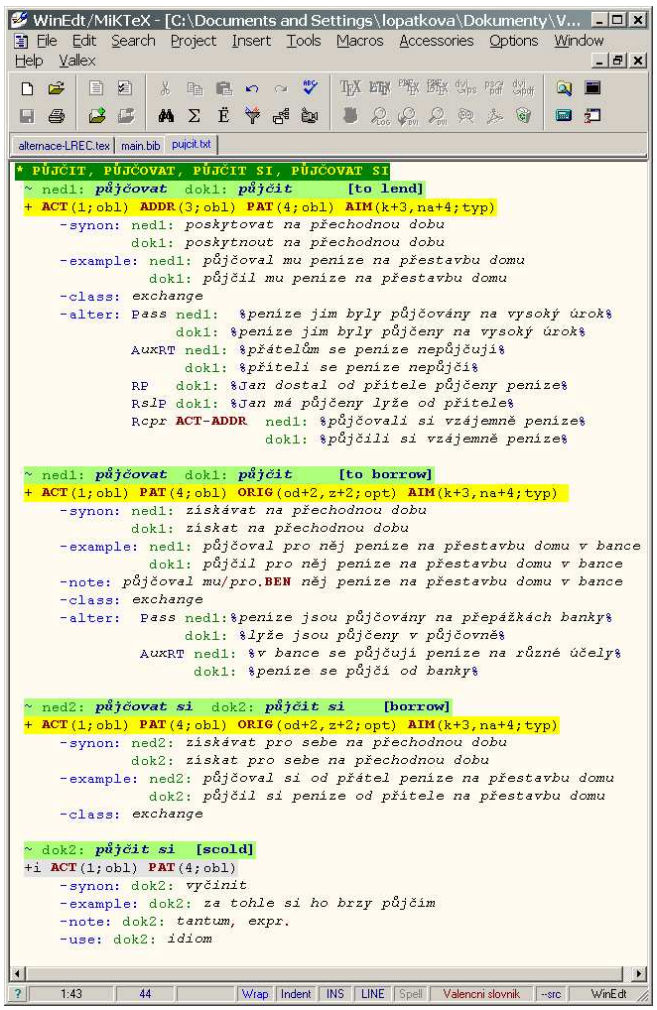


Figure 2: VALLEX lexeme for the lemma *půjčit/půjčovat/půjčit si/půjčovat si* [to lend / to borrow]. The highlighting mode in WinEdt text editor, the annotation tool for VALLEX.

The problem is that many verbs can be used in different contexts in the same or only slightly different meanings, which can be accompanied by small changes in their syntactic properties. When describing valency really explicitly, such changes imply introduction of new LUs, which is rather unintuitive and causes problems in building a lexicon (it is a substantial source of inconsistency during annotation, it causes redundancy in the lexicon). As an illustration:

(1) *Martin.ACT nastříkal barvu.PAT na zed'.DIR3*
Martin sprayed paint on the wall.

(2) *Martin.ACT nastříkal zed'.PAT barvou.*MEANS
Martin sprayed the wall with paint.

Clearly, different frames (containing different functors, i.e. labels of 'deep roles')[3] are instantiated in both pairs. Thus we have to have two LUs for these two utterances of verb

[3]Here the labels ACT and PAT stand for inner participants Actor/Bearer and Patient, respectively, the labels DIR3 and MEANS stand for free modifications Direction-where and Means.

despite the similarity of their meanings. The point here is that instead of having two unrelated LUs in the lexicon, it is more economical (less redundant) to store only one of them (considered as a basic LU) accompanied with information about particular alternation(s) that is/are applicable on this LU (and a derived LU can be generated 'on demand').

2.1. Threefold effect of alternations

In our approach, alternations are seen as transformations that take one LU as an argument and return another LU as a result. The effect of alternations is manifested by (at least one of) the following ways:

- change in **(complex) verb form**,
- change in **valency frame**, i.e.
  - changes in list of valency members,
  - changes in obligatorness of particular members,
  - changes in the sets of possible morphological forms of particular complementations,
- change in **lexical meaning** (with a possible change in the syntactic-semantic class).

Each alternation should be applicable on a whole group of LUs and its manifestation must be completely regular – all the changes (in form, in valency frame as well as in meaning) must be predictable from the input LU and the type of alternation.

2.2. Alternations as transformations

According to the alternation-based model, LUs are grouped into **LU clusters**, as is sketched in Fig. 3. Each cluster contains a **basic LU**, which has to be physically stored in the lexicon, and possibly a number of **derived LUs**, which are present only virtually in the lexicon – these derived LUs are obtained as results of transformations (for alternations applicable on the basic LU).

As the effects of alternations are completely regular, each alternation can be described in the grammar component of the lexicon as **set(s) of transformation rules** that can be applied on a basic LU. These transformations cover all changes in a LU relevant for a particular alternation.

Let us stress here that some alternations can be composed. Thus the LU cluster (see Fig. 3) can be seen as an oriented graph with one distinguished node (basic LU), from which there is an oriented path to all remaining nodes.

Concerning the choice of the basic LU, linguists do not offer in general any simple and explicit solution. Practically, this choice depends on the list of alternations introduced in the lexicon, so it is arbitrary to some extent (only the formal criterion that all other LUs are reachable from the chosen one must be fulfilled). Therefore certain conventions were adopted, some of them more obvious (as e.g. active construction is considered as the basic structure and particular passive constructions as the derived ones), other more arbitrary (as e.g. choice of basic LU for 'cause co-occurrence' alternation, see examples (5)-(6)).

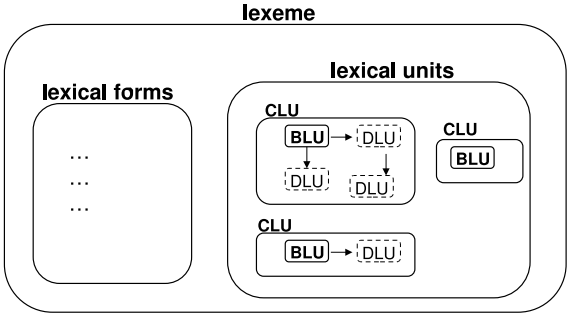


Figure 3: Basic and derived LUs (BLUs and DLUs) forming clusters of LUs (CLU).

Since some alternations can be combined the transformation rules must specify also changes in the list of alternations applicable to the output LU (see below, examples (3)-(4) and (5)-(6)).

The concept of transformations is described in detail on the 'recipient passive' alternation and 'cause co-occurrence' alternation in the following sections.

2.2.1. 'Recipient passive' alternation

The 'recipient passive' alternation can be exemplified on the sentences (3)-(4).

(3) *Pojišťovna.ACT zaplatila výrobcům.ADDR ztráty.PAT*
[insurance_company$_{Nom}$-covered-(to)producers$_{Dat}$-losses$_{Acc}$]
The insurance company covered losses to the producers.

(4) *Výrobci.ADDR dostali od pojišťovny.ACT zaplaceny ztráty.PAT*
[producers$_{Nom}$-got-from-insurance_company$_{Gen}$-covered-losses$_{Acc}$]
The producers have got covered their losses from the insurance company.

The active construction of a meaningful verb (here the verb *zaplatit* [to cover / to pay]) is considered as the basic LU, and thus it is contained in the VALLEX lexicon, see LU in Fig. 4. The set of applicable alternations (together with the examples) is listed in the atribute 'alter'.

It is specified in the grammar component, that the 'recipient passive' construction (marked RP in VALLEX) consists of the finite form of the verb *dostat* [to get] plus passive participle of the meaningful verb. The passive participle has either the form for neuter gender, or it agrees with the noun in accusative case (we draw on the description proposed in (Daneš, 1985) and (Mlu, 1987)).

Clearly, the 'recipient passive' construction has the same valency frame (i.e. the same set of valency complementations) as the active construction. However, the possible morphological forms are different – in active sentence, ACTor is in Nominative and ADDRessee in Dative case; in recipient passive, ACTor is either in Instrumental, or it is realized as a prepositional group *od* [from]+Genitive and ADDRessee is in Nominative (PATient is in Accusative case in both sentences).

ZAPLATIT
∼ pf: *zaplatit* [to cover / to pay]
+ ACT(1;obl) ADDR(3;opt) PAT(4;obl)
-gloss: *uhradit* [to cover / to pay]
-example: *zaplatit mechanikovi opravu*
[to pay the repair to a mechanic]
-class: *exchange*
-alter: Pass *%oprava byla zaplacena v eurech%*
[the repair was paid in euros]
AuxRT *%oprava se zaplatila v eurech%*
[the repair was paid in euros]
RP *%opravu dostali zaplacenu v eurech%*
[they have got the repair covered in euros]
RslP *%rodiče měli dovolenou zaplacenu %*
[parents have the holidays paid]
Rcpr ACT-ADDR
%zaplatili si (navzájem) všechny pohledávky%
[they covered their claims each to other]

Figure 4: The basic LU for the particular sense of the verb *zaplatit* [to cover / to pay] in the annotation format.

In VALLEX, a transformation notation developed by Petr Pajas (originally used for consistency checking of valency frames in PDT) was adopted for describing different types of alternations. Informally, the set of rules for RP alternation looks as follows:

- change in verb form:
  ⇒ +*dostat* [to get], finite form
  active ⇒ passive participle
  (neuter gender | agreement with the noun in Accusative)
- changes in valency frame :
  not applicable (NA in the sequel)
- changes of possible morphological forms:
  ACT(1) ⇒ – ACT(1), +ACT(7), +ACT(od+2)
  ADDR(3) ⇒ – ADDR(3), +ADDR(1)[4]
- change of syntactic-semantic class:
  NA
- change in the list of applicable alternations:
  ⇒ – Pass
  ⇒ – AuxRT
  ⇒ – RP
  ⇒ – RslP
  ⇒ – Rcpr

As a result of this transformation rule (applied to the basic LU for the verb *zaplatit* [to cover / to pay]), the derived LU for the 'recipient passive' construction is obtained, see Fig. 5 (the example is copied from the relevant alter attribute of the basic LU).

2.2.2. 'Cause co-occurrence' alternation

The 'cause co-occurrence' alternation concerns a group of verbs that express putting things / substances into containers or putting them on surface (for Czech described in (Daneš, 1985), for English see (Levin, 1993), Section 2.3).

[4]This is interpreted as: concerning ACT, remove Nominative case, add Instrumenal and prepositional group *od*+Genitive; concerning ADDR, remove Dative case and add Nominative.

∼ pf: *zaplatit* [to cover / to pay]
+ ACT(7,od+2;obl) ADDR(1;opt) PAT(4;obl)
-gloss: *uhradit* [to cover / to pay]
-example: *opravu dostali zaplacenu v eurech*
[they have got the repair covered in euros]
-class: *exchange*

Figure 5: The derived LU for the 'recipient passive' construction for the verb *zaplatit* [to cover / to pay].

(5) *Dělníci.ACT* naložili vagony.PAT uhlím.MEANS
The workers loaded the wagons with coal.

(6) *Dělníci.ACT naložili uhlí.PAT do vagonů.DIR3*
The workers loaded coal on the wagons.

Sentences (5)-(6) show two possible underlying syntactic structures that these verbs can create, see Table 1.

	agens / causator	container / surface	thing / substance
ex. (5)	ACT	PAT	MEANS
ex. (6)	ACT	DIR3	PAT

Table 1: Two possible underlying syntactic structures for the 'cause co-occurrence' alternation.

In VALLEX, the syntactic structure realized in the sentence (5) is considered as the primary one – thus the basic LU for the relevant sense of the verb *nakládat / naložit* [to load] is such as in Fig. 6 ('CCo' labels 'cause co-occurrence' alternation). All alternations applicable to this verb sense are presented here just to illustrate the possibility of alternations to compose.

NAKLÁDAT, NALOŽIT
∼ impf: *nakládat* pf: *naložit* [to load]
+ ACT(1;obl) PAT(4;obl) MEANS(;typ)
-gloss: impf: *plnit* pf: *naplnit* [to load]
-example: impf: *nakládat vůz senem*
pf: *naložit vůz senem*
[to load a wagon with hay]
-class: *providing*
-alter:
Pass impf: *%vozy byly nakládány dřevem po okraj%*
pf: *%vozy byly naloženy dřevem po okraj%*
[wagons were loaded with timber to the brim]
AuxRT impf: *%vozy se nakládaly dřevem po okraj%*
pf: *%vozy se naložily dřevem po okraj%*
[wagons were loaded with timber to the brim]
RslP pf: *%mít vůz naložený dřevem po okraj%*
[to have wagon loaded with timber to the brim]
CCo impf: *%nakládat seno na vůz%*
pf: *%naložit seno na vůz%*
[to load hay on wagon]

Figure 6: The basic LU for the particular sense of the verb *nakládat / naložit* [to load].

The transformation rule in the grammar component of VALLEX specifies the way how to obtain a derived LU for particular alternations. Concerning CCo, the following changes are relevant:

• change in verb form:
  NA

• changes in valency frame (list of complementations as well as obligatorness of particular members):
  MEANS ⇒ – MEANS
  ⇒ +DIR3(;obl)

• changes of possible morphological forms:
  NA

• change of syntactic-semantic class:
  providing ⇒ location

• change in list of applicable alternations:
  ⇒ – CCo

The result of the CCo transformation rule applied to the appropriate basic LU for the verb *nakládat / naložit* [to load] is shown in Fig. 7.

NAKLÁDAT, NALOŽIT
∼ impf: *nakládat* pf: *naložit* [to load]
+ ACT(1;obl) PAT(4;obl) DIR3(;obl)
-gloss: impf: *plnit* pf: *naplnit* [to load]
-example: impf: *nakládat seno na vůz*
pf: *naložit seno na vůz*
[to load hay on wagon]
-class: *location*
-alter: Pass
AuxRT
RslP

Figure 7: The derived LU for the 'cause co-occurrence' alternation for the verb *nakládat / naložit* [to load].

As the lists of alternations applicable to derived LU's are gained from the transformation rules in the grammar component (not from the data component), there cannot be examples of their instantiations in derived LUs (we minimize this minus by ordering alternations, see Section 2.3.).

2.3. Typology of alternations

Basically, we distinguish two groups of alternations, tentatively characterized as 'syntactically-based' alternations and 'semantically-based' ones.

2.3.1. 'Syntactically-based' alternations

A group of 'syntactically-based' alternations primarily consists of different types of 'diathesis' (in the narrow sense) in Czech. Further, reciprocal alternations are ranged with this type and also some additional (more sparse) constructions. These alternations are characterized by changes in the verb form.

We have exemplified some of these alternations in the previous section in Figures 4 and 6, where label Pass stands for passive voice, AuxRT for reflexive passive, RP and RslP for recipient and resultative passive with *dostat* [to get] and *mít* [to have], respectively, plus passive participle constructions. We take into account also, e.g., alternations for constructions like *dát / nechat* plus infinitive (as in *dává / nechává si vyprat špinavé košile* [he has/gets his dirty shirts washed]). Label Rcpr (see Fig. 4) is used for reciprocal constructions described for Czech in (Panevová, 1999).

The 'syntactically-based' alternations cover constructions described in details in Czech grammars, another 'diatheses' are regular enough to be covered by general rules (e.g. 'dispositional modality' or impersonal constructions), so it is redundant to store them in a lexicon (see esp. (Mlu, 1987) and (Daneš, 1985), and (Skoumalová, 2002)).

2.3.2. 'Semantically-based' alternations

Let us give here at least several examples to illustrate 'semantically-based' alternations. Levin stated that alternations are language dependent, though several of English examples have their Czech counterparts, e.g. 'cause co-occurrence' alternation (examples (1)-(2)) matches up with Levin's 2.3 alternations (see also (Cinková, 2006)). The following Table 2 shows some other examples of semantically-based alternations (examples marked with ⋆ are described in (Benešová, 2004)).

1.4	*vyjít kopec / vyjít na kopec*⋆ [to climb the mountain / to climb up the mountain]
2.4	*chlapec roste v muže / z chlapce roste muž* [a boy grows into a man / a man grows from a boy]
1.1	*Slunce vyzařuje teplo / teplo vyzařuje ze slunce* [the Sun radiates heat / heat radiates from the Sun]
2.1	*poslat dopis mamince / poslat peníze do Indie*⋆ [to send mamma a letter / to send money to India]
???	*soustředit se v centru města/ soustředit se do centra*⋆ [to mass in the city center / to mass into the city center]

Table 2: Examples of corresponding Czech and English alternations (numbers in first column stand for Levin's types of alternations).

Distinguishing two basic groups of alternations is not an enterprize for its own sake – these two groups exhibit different behavior:

• Alternations belonging to the same group typically cannot be composed (with the rare exception of Rcpr alternation where subject is not involved – this case must be treated separately).

• Typically, alternations from different groups can be mutually composed.

• Though in general, alternations from different groups can be composed in any order, we have not found a single example where the order of composition is relevant. That means that the result of composition is the same regardless the order.

These observations result in an important constraint – it allows us to prescribe the order in which alternations can be composed: if two alternations are to be applied to any LU, then the 'semantically-based' one is (by convention) considered as the first one, the 'syntactically-based' one follows.

This constraint has both theoretical and practical impact. It guarantees the tree structure of LU clusters (compare Fig. 3 in Section 2.). From the practical point of view it ensures that 'semantically-based' alternations are exemplified in the

lexicon. Considering the exhaustive description of passive constructions in grammar books (and also description of other constructions which come under 'syntactically-based' alternations), it seems to be acceptable to have these types of alternations without examples in the expanded form of the lexicon.

3. Minimal and expanded form of the lexicon

The VALLEX lexicon (in its minimal form) contains only the basic LU with an associated list of applicable alternations. However, there are various tasks for which it could be useful to include the derived LUs to the lexicon (e.g. frame disambiguation, i.e. assigning LUs to verb occurrences in text). This requirement leads to distinguishing minimal and expanded form of valency lexicon VALLEX – the expanded one (containing all LUs covered either explicitly or implicitly in the lexicon) can be derived from the minimal one (containing only basic LUs) by a fully automatic procedure. The formal alternation-based model of VALLEX is described in details in (Žabokrtský, 2005), where also the main software components of the dictionary production system developed for VALLEX are outlined (including annotation format, www interface for searching the text format as well as XML data format).

Conclusions

Despite the variety of valency behavior of lexical units, in the valency lexicon of Czech verbs VALLEX the stress is laid on an adequate and consistent description of regular properties of verbs as lexical units. The alternation-based model gives a more powerful description of Czech verbs and shows regular changes in their argument structure. It makes it possible to decrease redundancy in the lexicon and to make the lexicon more consistent.

In future, we will especially focus on the 'semantically-based' alternations in Czech, the adequate description of which requires further linguistic research. We aim to empirically confirm the adequacy of tree-structure constraint on LU clusters. Depending on the progress in this field, we intend to involve newly specified alternations to the lexicon. We plan to extend VALLEX also in quantitative aspects.

The alternation-based model is a novelty in Czech computational lexicography. Though only a limited number of alternations has been practically implemented in VALLEX, its asset to adequate description of valency properties of verbs has been clearly proved.

Acknowledgement

The research reported in this paper has been supported by the grant of the Grant Agency of Czech Republic No. 405/04/0243.

4. References

Olga Babko-Malaya, Martha Palmer, Nianwen Xue, Aravind Joshi, and Seth Kulick. 2004. Proposition Bank II: Delving Deeper. In A. Meyers, editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 17–23, Boston, USA. ACL.

Václava Benešová. 2004. Delimitace lexií českých sloves z hlediska jejich syntaktických vlastností. Master's thesis, Filozofická fakulta Univerzity Karlovy.

Silvie Cinková. 2006. From PropBank to EngValLex. In *Proceedings of LREC 2006*. (this volume).

D. A. Cruse. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge.

František Daneš. 1985. *Věta a text*. Academia, Praha.

Katrin Erk, Andrea Kowalski, Sebastian Pado, and Manfred Pinkal. 2003. Towards a Resource for Lexical Semantics: A Large German Corpus with Extensive Semantic Annotation. In *Proceedings of ACL-03*, Sapporo, Japan.

Josef Filipec. 1994. Lexicology and Lexicography: Development and State of the Research. In Philip A. Luelsdorff, editor, *The Prague School of Structural and Functional Linguistics*, pages 164–183. John Benjamins Publishing Company.

Charles J. Fillmore. 1968. The Case for Case. In E. Bach and R. Harms, editors, *Universals in Linguistic Theory*, pages 1–90. New York.

Charles J. Fillmore. 2002. FrameNet and the Linking between Semantic and Syntactic Relations. In Shu-Cuan Tseng, editor, *Proceedings of COLING 2002*, pages xxviii–xxxvi. Howard International House.

Jan Hajič, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová, and Petr Pajas. 2003. PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9, pages 57–68.

Jan Hajič. 2005. Complex Corpus Annotation: The Prague Dependency Treebank. In *Insight into Slovak and Czech Corpus Linguistics*, pages 54–73. Veda Bratislava, Slovakia.

Beth C. Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago, IL.

Markéta Lopatková. 2003. Valency in Prague Dependency Treebank: Building Valency Lexicon. *The Prague Bulletin of Mathematical Linguistics 79-80*.

Igor A. Mel'čuk and Alexander K. Zholkovsky. 1984. *Explanatory Combinatorial Dictionary of Modern Russian*. Wiener Slawistischer Almanach, Vienna.

1987. *Mluvnice češtiny III*. Academia, Praha.

Jarmila Panevová. 1994. Valency Frames and the Meaning of the Sentence. In Philip A. Luelsdorff, editor, *The Prague School of Structural and Functional Linguistics*, pages 223–243. John Benjamins Publishing Company.

Jarmila Panevová. 1999. Česká recipročnı́ zájmena a slovesná valence. *Slovo a slovesnost*, 4(60):269–275.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company, Dordrecht.

Hana Skoumalová. 2002. Verb frames extracted from dictionaries. *The Prague Bulletin of Mathematical Linguistics 77*.

Zdeněk Žabokrtský. 2005. *Valency Lexicon of Czech Verbs*. Ph.D. thesis, Charles University in Prague, Faculty of Mathematics and Physics.

B.3 Changes in Valency Structure of Verbs: Grammar vs. Lexicon

In *Proceedings of Slovko 2009, NLP, Corpus Linguistics, Corpus Based Grammar Research*, p. 198-210, Bratislava, Slovakia, 2009. Slovenská akadémia vied
(with co-author V. Kettnerová)

Changes in Valency Structure of Verbs: Grammar vs. Lexicon*

Václava Kettnerová and Markéta Lopatková

Institute of Formal and Applied Linguistics
Charles University in Prague, Czech Republic

Abstract. In this paper, we deal with changes in valency structure of Czech verbs from a lexicographic point of view. We focus only on syntactic constructions that are related in principle to the same (generalized) situation. Changes in valency structure are understood as different mappings between individual participants of a generalized situation and valency slots, including their morphemic realization. We distinguish two types of changes in valency structure, so-called grammatical diatheses and semantic diatheses. We introduce a basic typology of potential changes in valency structure and we propose a method of the representation of these changes in the valency lexicon of Czech verbs VALLEX.

1 Motivation

Syntactic behavior of verbs is determined to a great extent by their lexical semantic properties. Prototypically, a single valency structure corresponds to a single meaning of verb. However, in many 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. *Peter loaded the truck with hay.* — b. *Peter loaded hay on the truck.*

(2) a. *The truck was loaded with hay.* — b. *Hay was loaded on the truck.*

Such uses of the verb *load* cannot be described by a single valency frame; however, separating four valency frames appears to be redundant with respect to the regularity in morphemic realizations of valency slots. Let us focus on the pairs of sentences (1a)-(2a) and (1b)-(2b). In these cases, (i) the information on the possibility of such change in valency structure of the verb *load* and (ii) the rule describing such change are sufficient for lexicographic description. Other changes in valency structure of verbs can be treated in a similar way under the condition that these changes are so regular that they can be captured by means of rules.

In this contribution, we deal with changes in valency structure of Czech verbs from a lexicographic point of view. We introduce and exemplify a basic typology of potential changes in valency structure of Czech verbs as they have appeared during the lexicographic processing language data (based on corpus evidence). Finally, we propose a method of representing these changes in a valency lexicon of Czech verbs.

* The research is carried under the MŠMT ČR project No. MSM0021620838 and partially under the MŠMT grant No. LC536 and GA UK grant No. 7982/2007.

Basic approaches to changes in valency structure. 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. [1, 3, 6]. The terms hierachization, diathesis or conversion were introduced in Czech and Slovak grammars, see esp. [7, 8, 15, 21] and [11]. Roughly speaking, such terms refer to change 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 three basic approaches to changes in valency structure of verbs, (i) structurally based approaches represented mainly by transformational-generative grammars, esp. [4, 5], (ii) lexically based approaches focusing on the relation between lexical semantic properties of verbs and their syntactic behavior, esp. [12], and (iii) constructionally based approaches based on the assumption that difference in syntactic forms marks the difference in meaning, esp. [2, 10].

Here we focus on the description of changes in valency structure of verbs in the theoretical framework of the Functional Generative Description (FGD), see esp. [20]. The valency theory of FGD, esp. [16], was applied to a large number of data in building the Prague Dependency Treebank, PDT 2.0[1] and the valency lexicon of Czech verbs, VALLEX[2] [13]. We attempt to propose an adequate framework for description of changes in valency structure of verbs which can be applied in lexicographic processing of language data.

2 Basic typology of changes in valency structure of verbs

In our typology of changes in valency structure of verbs, the concept of situation plays a key role. The **(generalized) situation** represents a class of abstract situations characterized by a particular set of semantic participants.[3] In the present paper, we focus only on those syntactic constructions that relate to the same (generalized) situation. Such a situation is expressed by a single verb lexeme and it is characterized by an identical set of semantic participants. Changes in valency structure are understood as different mappings between individual semantic participants of a generalized situation and their surface syntactic positions, including their morphemic realization. We distinguish two types of changes in valency structure, so-called grammatical diatheses (g-diatheses) and semantic diatheses (s-diatheses).

2.1 Grammatical diatheses

G-diatheses represent pairs of related syntactic constructions that prototypically satisfy the following criteria:

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

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

[3] See also type situation [8, 22] or semantic event. Semantic participants roughly correspond to semantic roles here.

I. Verbs in the marked construction are prototypically morphologically marked with respect to the grammatical category of voice. Their forms typically either consist of auxiliaries and non-finite form of lexical verbs or they have reflexive forms.
II. The mapping between semantic participants of a generalized situation and valency slots remains unchanged, their number and type are identical as well. Changes in valency frames are typically connected with a choice of a particular valency member for the subject syntactic positions; these changes are limited to morphemic realizations of individual valency slots.

G-diatheses primarily represent a language means that enables the speaker to choose a particular semantic participant of a generalized situation for the syntactically prominent position of (surface) subject. In the marked case, the valency member ACT (Actor, corresponding to the semantic participants of generalized situation such as Agent, Initiator, Causator, Bearer of Action, etc.) is prototypically shifted from the subject syntactic position into a less prominent surface position; eventually, it cannot be expressed on the surface syntactic level at all (as in deagentive g-diathesis, see e.g. [9]). Another semantic participant of a generalized situation (typically having the form of accusative) is shifted into the subject syntactic position, as in (1a)-(2a) repeated below.[4] Under certain conditions, a 'subject-less' construction occurs (see example (7b) below).

(1) a. *Peter*.ACT *loaded the truck*.PAT *with hay*.EFF

(2) a. *The truck*.PAT *was loaded with hay*.EFF (*by Peter*.ACT)

G-diatheses can be illustrated by the scheme in Figure 1, the asymmetry concerns the different mappings between a set of valency members and their surface positions.

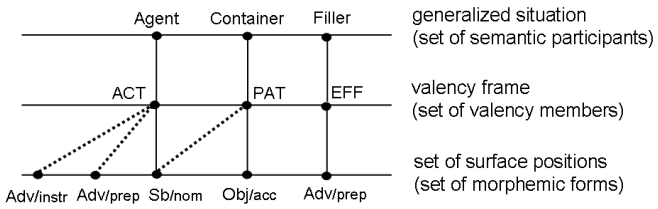


Fig. 1. Mapping between semantic participants of a generalized situation and their surface syntactic positions for passive diathesis as a typical g-diathesis (for the verb *naložit* 'to load').

We assume that changes in the valency structure of verbs characteristic of g-diatheses arise from the special verbal meanings. These verbal meanings are reflected as values of relevant verbal grammatemes in FGD (grammatemes represent tectogrammatical correlates of the morphological categories, see [14, 19]).

[4] We mark the valency members with labels (so-called functors) ACT, PAT, EFF etc. in accordance with FGD (and with VALLEX in particular).

2.2 Semantic diatheses

S-diatheses are characterized by changes in number and type of valency slots, while the (generalized) situation still remains unchanged. Furthermore, verbs are not morphologically marked with regard to voice. Contrary to g-diatheses, it is not apparent which of the related constructions should be understood as unmarked ones and which as marked ones, see also [8].

Moreover, s-diatheses are typically associated with coherent semantic classes of verbs, as in sentences (1a)-(1b) (see also, e.g., *spray/load* verbs in [12]).

(1) a. *Peter*.ACT-Agent *loaded the truck*.PAT-Container
with hay.EFF-Filler
b. *Peter*.ACT-Agent *loaded hay*.PAT-Filler
on the truck.DIR-Container

In Czech grammars, s-diatheses are described as hierarchizations without marked voice [8], as objective diatheses [11], or some of them are treated as examples of the so-called decauzativization [11].

S-diatheses can be illustrated by the scheme in Figure 2, the asymmetry concerns the different mappings between a set of semantic participants of a generalized situation and a set of valency members.

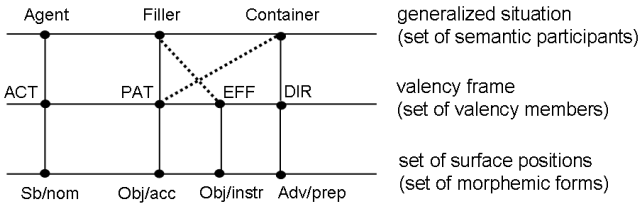


Fig. 2. Mapping between semantic participants of a generalized situation and their surface syntactic positions for Container-Filler diathesis (for the verb *naložit* ‘to load’).

As to the possibility of combining g- and s-diatheses, diatheses of different types are mutually combinable; i.e., having a marked construction with respect to a g-diathesis, a particular s-diathesis rule may be subsequently used (if applicable for the given verb), and conversely, see ex. (1)-(2) in Section 1. However, mutually combining diatheses of the same type is very restricted.[5]

Distinguishing between g-diatheses and s-diatheses is motivated by the needs of lexicographic work. We will see later that in case of **g-diatheses**, the changes in valency

[5] E.g., *Když se dostane přidělena pracovna, to se to pracuje.* — Eng. If a new study is allocated, it is easy to work (example from [9]).

frames are regular enough to be treated within a single verbal lexical unit – general rules in the grammar component and information on their applicability to individual lexical units in the data component of the lexicon are sufficient. However, for **s-diatheses**, we propose to set separate lexical units interlinked with general rules identifying a relevant type of s-diathesis. This solution results from the corpus evidence that changes in valency structure of verbs are diverse even within an individual type of s-diatheses.

3 Representation of G-diatheses

In this section, we introduce a way of capturing g-diatheses in the valency lexicon VALLEX. In our approach, g-diatheses are described by means of general fine-grained rules in the grammar component of the valency lexicon. All applicable g-diatheses are listed for each verbal lexical unit separately in a special attribute in the data component of the lexicon.

Our method will be demonstrated on the passive diathesis as a prototypical g-diathesis. **Deagentive diathesis**, **recipient diathesis**, **resultative diathesis** and **mediopassive diathesis**, see esp. [19], can be described in the same way. In addition, we consider also **reciprocity** as a phenomenon that can be treated in a similar way (within FGD, reciprocity and the possibility of its representation have been broadly studied by Panevová, esp. [17]).[6]

3.1 Passive diathesis

Passive diathesis is a relation between two syntactic constructions in which the marked one contains the auxiliary verb *být* 'to be' and the past participle of a lexical verb. We propose the following representation of passive diathesis in the valency lexicon:

(i) In the **data component**, a single lexical unit is represented by an (unmarked) valency frame. If a given lexical unit can be subject to passive diathesis, then its applicability is indicated in the special attribute 'diathesis-pass'.
(i) In the **grammar component**, a general rule describing regular changes in a valency frame for this diathesis is stored.

For example, a lexical unit for the transitive verb *postavit* 'to build' has three valency slots in its valency frame: obligatory ACT (Actor, in nominative in the unmarked construction), obligatory PAT (Patient, in accusative) and optional ORIG (Origin, expressed as the prepositional group *z* 'from/of' plus genitive). In the marked construction, ACT is realized either as instrumental or as prepositional group *od* 'by' plus genitive, and PAT is expressed as nominative (morphemic realization of ORIG remains unchanged):

(3) a. *David*.ACT$_{nom}$ *postavil kůlnu*.PAT$_{acc}$ *ze dřeva*.ORIG$_{z+gen}$
Eng. David.ACT built a shed.PAT from wood.ORIG
b. *Kůlna*.PAT$_{nom}$ *byla postavena ze dřeva*.ORIG$_{z+gen}$ (*Davidem / od Davida*.ACT$_{instr,od+gen}$)
Eng. A shed.PAT was built from wood.ORIG (by David.ACT)

[6] **Causative constructions** are another candidates that can be taken into account for this type of representation.

Passive diathesis for verbs with valency member expressed by accusative. Passive diathesis concerns verbs with at least two semantic participants of a generalized situation and thus at least two valency slots, prototypically ACT in nominative and PAT in accusative. Valency frame for the marked member of the diathesis can be described by the following rule **Pass.r1.PAT**, see Table 1.

It should be stressed here that all information captured in valency frame remains unchanged, unless a change is explicitly mentioned by the rule Pass.r1.PAT; i.e., if a valency frame contains a member or morphemic form that is not cited in the rule, then it is preserved also in a derived valency frame.

Pass.r1.PAT	Unmarked	Marked	Note
verbal grammateme	diathesis-pass: 0	diathesis-pass: pass	(1)
valency frame	ACT$_{nom}$	ACT$_{instr,od+gen}$	(2)
	PAT$_{acc}$	PAT$_{nom}$	(3)
	PAT$_{var,inf,dcc}$	PAT$_{excluded}$	(4)
	? EFF$_{jako+acc}$	? EFF$_{jako+nom}$	(5)

Table 1. Pass.r1.PAT rule for the passive diathesis.

Commentary on the Pass.r1.PAT rule:
(1) The passive diathesis is represented by the verbal grammateme 'diathesis-pass'; its value for the unmarked member of the pair is '0', for the marked member it is 'pass'.
(2) In the marked construction, ACT is shifted from the prominent subject syntactic position into the adverbial position. This change is accompanied by the change of morphemic realization of ACT from nominative into instrumental or into the prepositional case *od* 'by'+genitive.
(3) The valency member PAT (expressed by accusative) is selected for the prominent surface syntactic position of subject for the marked member of the passive diathesis. Its morphemic form is changed into nominative.
(4) If the PAT valency member may be expressed also by other morphemic forms such as infinitive (abbr. *inf*), dependent content clause (*dcc*) or another preposition or prepositionless case (*var*) (mentioned below as 'unaccusative variants'), all these possible morphemic variants are excluded in the marked frame. PAT expressed by unaccusative forms is treated with Pass.r2.PAT rule, see below.
(5) If there is a slot for EFF in the unmarked frame with the form *jako* 'as'+accusative, then its form is changed into *jako* 'as'+nominative.
Note on agreement: Verbal categories of person, number and gender agree with ACT in nominative in the unmarked construction, whereas a verb in the marked construction has agreement with PAT in nominative.

For example, by applying Pass.r1.PAT rule to the unmarked valency frame for the verb *postavit* 'to build', see ex. (3a)-(3b), we obtain the following valency frame describing the marked syntactic construction:

ACT$_{nom}$ PAT$_{acc}$ ORIG$_{z+gen}$
$\Rightarrow_{\text{Pass.r1.PAT}}$ ACT$_{instr,od+gen}$ PAT$_{nom}$ ORIG$_{z+gen}$

The change in the realization of EFF expressed with *jako* 'as'+accusative may be exemplified by the verb *hodnotit* 'to assess'. See the unmarked and marked valency frames and their realizations in sentences (4a)-(4b) (note also the reduction of possible morphemic forms for PAT in (4b)):

$\text{ACT}_{nom}\ \text{PAT}_{acc,var,inf,dcc}\ \text{EFF}_{jako+acc,na+acc}$ $\Rightarrow_{\text{Pass.r1.PAT}}\ \text{ACT}_{instr,od+gen}\ \text{PAT}_{nom}\ \text{EFF}_{jako+nom,na+acc}$

(4) a. *Učitelé*.ACT$_{nom}$ *hodnotili jeho práci*.PAT$_{acc}$ *jako nedostatečnou*.EFF$_{as+acc}$
Eng. The teachers.ACT assessed his paper.PAT as poor.EFF
b. *Jeho práce*.PAT$_{nom}$ *byla hodnocena učiteli*.ACT$_{instr}$ *jako nedostatečná*.EFF$_{as+nom}$
Eng. His paper.PAT was assessed as poor.EFF by his teachers.ACT

For some verbs with at least three valency members, the accusative position may be labeled with other functors, namely ADDR (for Addressee) or EFF (for Effect),[7] see (5a)-(5b) and (6a)-(6b). The changes in valency structure of these verbs are captured by analogous rules Pass.r1.ADDR and Pass.r1.EFF.

(5) a. *Sekretářka*.ACT$_{nom}$ *ředitele*.ADDR$_{acc}$ *upozornila,* (*že má podepsat smlouvu*).PAT$_{dcc}$
Eng. The secretary.ACT has reminded the director.ADDR (to sign the contract).PAT
b. *Ředitel*.ADDR$_{nom}$ *byl upozorněn sekretářkou*.ACT$_{instr}$, (*že má podepsat smlouvu*).PAT$_{dcc}$
Eng. The director.ADDR has been reminded by his secretary.ACT (to sign the contract).PAT

(6) a. *Zadržený*.ACT$_{nom}$ *řekl vyšetřovateli*.ADDR$_{dat}$ *lež*.EFF$_{acc}$
Eng. The detained man.ACT said to the interrogator.ADDR a lie.EFF
b. *Vyšetřovateli*.ADDR$_{dat}$ *byla* (*zadrženým*.ACT$_{instr}$) *řečena lež*.EFF$_{nom}$
Eng. A lie.EFF was said to the interrogator.ADDR (by the detained man.ACT)

Passive diathesis for verbs with valency member expressed by 'unaccusative' forms. Furthermore, passive diathesis can be applied to verbs with valency members realized by 'unaccusative' forms, see ex. (7a)-(7b):

(7) a. *Radní*.ACT$_{nom}$ *o té záležitosti*.PAT$_{o+loc}$ *rozhodli včera.*
Eng. The councilors.ACT decided the matter.PAT yesterday.
b. *O té záležitosti*.PAT$_{o+loc}$ *bylo* (*radními*.ACT$_{instr}$) *rozhodnuto včera.*
Eng. The matter.PAT was decided (by councilors.ACT) yesterday.

Changes in valency frame are described by the following rule **Pass.r2.PAT**, see Table 2. Again, except for the changes explicitly mentioned in the rule, all other information captured in a valency frame remains unchanged.

[7] We leave aside the functors DPHR (for Dependent Part of Phraseme) and CPHR (Part of Compound Predicate) here.

Pass.r2.PAT	Unmarked	Marked	Note
verbal grammateme	diathesis-pass: 0	diathesis-pass: pass	(1)
valency frame	ACT_{nom}	$\text{ACT}_{instr,od+gen}$	(2)
	$\text{PAT}_{var,inf,dcc}$	$\text{PAT}_{var,inf,dcc}$	(3)
	? PAT|ADDR|EFF_{acc}	? PAT|ADDR|$\text{EFF}_{excluded}$	(4)

Table 2. Pass.r2.PAT rule for the passive diathesis.

Commentary on the Pass.r2.PAT rule:
(1) and (2) See the Commentary on the Pass.r1 rule.
(3) The 'unaccusative' morphemic realization of PAT[8] remains unchanged. If PAT is realized by infinitive or dependent content clause, it is shifted into the subject syntactic position. Applying the given rule to PAT expressed by prepositional case or prepositionless case (with the exception of accusative), 'subject-less' sentence is created.
(4) The possible accusative realization of any valency slot is excluded. If no other morphemic variant remains, the given valency member cannot be realized in a surface sentence,[9] see also ex. (8c).
Note on agreement: In the marked construction, verbs have incongruent agreement with 3^{rd} sg. neutr.

Let us exemplify the application of Pass.r2.PAT rule to the valency frame of the verb *rozhodnout* 'to decide', see also sentences (7a)-(7b):

$$\text{ACT}_{nom}\ \text{PAT}_{o+loc,dcc} \Rightarrow_{\text{Pass.r2.PAT}} \text{ACT}_{instr}\ \text{PAT}_{o+loc,dcc}$$

Verbs allowing for two passive constructions. There are verbs allowing for two passive constructions. First, such verb has a valency member that may be realized both as accusative and 'unaccusative' form (e.g., the verb *hodnotit* 'to asses', see ex. (4)) – then both types of rules are applicable to this valency member (Pass.r1.PAT or Pass.r2.PAT for the verb *hodnotit* 'to asses'). The second case is represented by verbs with at least three semantic participants of generalized situations. Such verbs have at least three valency members (prototypically realized as nominative, accusative and 'unaccusative').[10] Again, both types of rules may be used – they are applied to two different valency members depending on the choice of subject. We exemplify this by the verb *žádat* 'to ask', see sentence (8a) for the unmarked case, (8b) for the Pass.r1.ADDR rule and (8c) for the Pass.r2.PAT rule:

$$\text{ACT}_{nom}\ \text{ADDR}_{acc}\ \text{PAT}_{o+acc,inf,dcc} \Rightarrow_{\text{Pass.r1.ADDR}} \text{ACT}_{instr,od+gen}\ \text{ADDR}_{nom}\ \text{PAT}_{o+acc,inf,dcc}$$

$$\text{ACT}_{nom}\ \text{ADDR}_{acc}\ \text{PAT}_{o+acc,inf,dcc} \Rightarrow_{\text{Pass.r2.PAT}} \text{ACT}_{instr,od+gen}\ \text{ADDR}_{general}\ \text{PAT}_{o+acc,inf,dcc}$$

[8] The analogous rules are set for ADDR and EFF.
[9] This case results in so called generalized valency member in FGD, see [18].
[10] The verb *učit* 'to teach' with two valency members expressed in accusative represents a rare exception.

As the accusative is the only possible realization of ADDR in the unmarked valency slot (and accusative is excluded in the marked valency frame according to Pass.r2.PAT rule), the ADDR valency slot cannot be realized in the surface sentence, see ex. (8c).

(8) a. *Novináři*.ACT$_{nom}$ *vládu*.ADDR$_{acc}$ *žádali,* (*aby byly zveřejněny výsledky*).PAT$_{dcc}$
Eng. The journalists.ACT asked the government.ADDR (to publish the results).PAT
b. *Vláda*.ADDR$_{nom}$ *byla* (*novináři*.ACT$_{instr}$) *žádána,* (*aby byly zveřejněny výsledky*).PAT$_{dcc}$
Eng. The government.ADDR was asked (by the journalists.ACT) (to publish the results).PAT
c. *Novináři*.ACT$_{instr}$ *bylo opakovaně žádáno,* (*aby byly zveřejněny výsledky*).PAT$_{dcc}$ (general ADDR)
'(by) journalists - was - repeatedly - asked - to - publish - results' Eng. The publication of the results was repeatedly asked (by the journalists).

4 Representation of S-diatheses

In this section, we focus on s-diatheses and their adequate representation in the valency lexicon VALLEX. To recapitulate, s-diathesis is a relation between two (or more) syntactic constructions describing a same generalized situation. These constructions refer to the same (polysemous) verb lexeme, however, the mappings between individual semantic participants of the generalized situation and valency slots is different. As a consequence, not only morphemic realization but also number, type and obligatoriness of valency members may differ. In contrast to g-diatheses, morphological categories of the given verb typically remain unchanged.

Let us demonstrate our approach on the Container-Filler diathesis as a prototypical s-diathesis. Other s-diatheses can be captured in the same way (selected examples are listed below).

4.1 Container-Filler diathesis

Container-Filler diathesis[11] can be exemplified by sentences (9a)-(9b) (note that 'negative' variant can be also distinguished).

(9) a. *Petr*.ACT$_{nom}$-Agent *naložil vůz*.PAT$_{acc}$-Container *senem*.EFF$_{instr}$-Filler
Eng. Petr.ACT-Agent loaded the truck.PAT-Container with hay.EFF-Filler
b. *Petr*.ACT$_{nom}$-Agent *naložil seno*.PAT$_{acc}$-Filler

[11] This type of diathesis counts among a group of 'co-occurrence diathesis' in [8]; see also 'spray/load alternation' in [12]. We adopt a labeling based on semantic participants involved in the diatheses as we consider it more transparent.

na vůz.DIR-Container
Eng. Petr.ACT-Agent loaded hay.PAT-Filler
on the truck.DIR-Container

These two sentences describe in principle the same generalized situation with three semantic participants – Agent (who causes the action described by the given verb), Filler (substance or entity whose location is changed) and Container (location where Filler is moved). Despite the single set of semantic participants of the generalized situation, this situation can be structured in a different way. While Agent is realized as ACT in both cases, there are two possibilities for Filler and Container: (i) either Container is mapped onto PAT (in accusative) and Filler is mapped onto EFF valency slot (in instrumental), as in (9a); (ii) or Filler occupies the PAT slot (in accusative) and Container is structured as Directional modification DIR, as in (9b) (see also Figure 2 in Section 2.2).

The most studied semantic property of this diathesis deals with a partitive / holistic effect. The semantic participant of the generalized situation realized as PAT in accusative typically receives holistic interpretation; i.e., in Container-Filler diathesis either Container (9a) or Filler (9b) is understood as completely affected by the action expressed by the verb *naložit* 'to load'.

Contrary to g-diatheses, the changes in valency frames accompanying s-diatheses are not regular enough: individual verbs exhibit many irregularities in their valency characteristics even within a single type of s-diathesis (see below for the examples).

For the purpose of the valency lexicon VALLEX, we propose the following representation of s-diatheses:

(i) In the **data component**, we establish a set of two lexical units within one lexeme – each member of s-diathesis is represented by a separate lexical unit with its own valency frame. These lexical units are interlinked via the type of s-diathesis (captured in a special attribute 's-diathesis').
(ii) In the **grammar component**, a general rule describing possible mappings between semantic participants of a generalized situation and individual valency slots is provided, see Table 3.

Container-Filler	**Agent**	**Filler**	**Container**	examples
Filler $\sim$ **PAT**	ACT	PAT	DIR	*naložit seno na vůz* *doplnit cukr do cukřenky* *nasypat mouku do pytle* *(na)točit vodu (do kýble)*
Container $\sim$ **PAT**	ACT	EFF	PAT	*naložit vůz senem* *doplnit cukřenku cukrem/o cukr*
	ACT	—	PAT	*nasypat pytel *moukou* *(na)točit kýbl *vodou*

Table 3. General rule for the Container-Filler diathesis (see the translations below).

The dissimilarities in the Container-Filler diathesis concern number, type, and morphemic realization of complements as well:

- Whereas the set of semantic participants of the generalized situation is the same (Agent, Filler, Container) and prototypically all of them can be realized as valency members, this does not hold for some verbs (e.g., *nasypat mouku do pytle* 'to put flour into the sack' but *nasypat pytel *moukou* 'to put the sack *with flour').
- Whereas directional valency member that realizes Container participant is prototypically obligatory (e.g., *doplnit cukr do cukřenky* 'to add sugar to the sugar bowl'), there are verbs with only typical directional valency member (e.g., *točit vodu (do kýble)* 'to draw water (to the bucket)').
- Morphemic realizations of a particular valency member may differ with individual verbs (e.g., *doplnit cukřenku cukrem / o cukr* 'to replenish the sugar bowl with sugar').

4.2 Examples of other S-diatheses

While g-diatheses are intensively studied in Czech linguistics, there is only a limited number of studies of phenomena referred here to as s-diatheses, see esp. [8]. Let us exemplify here at least several frequent s-diatheses in Czech which can be captured in the valency lexicon in a similar way as the Container-Filler diathesis:

Surface-Cover diathesis (positive or negative)
Jana si očistila bláto.PAT-Cover *z bot*.DIR-Surface
Eng. Jane cleaned the mud.PAT-Cover off her shoes.DIR-Surface
— *Jana si očistila boty*.PAT-Surface *od bláta*.ORIG-Cover
Eng. Jane cleaned her shoes.PAT-Surface of the mud.ORIG-Cover

Material-Product diathesis (positive or negative)
Kadeřník jí učesal vlasy.PAT-Material *do drdolu*.EFF-Product
Eng. Hairdresser arranged her hair.PAT-Material into a bun.EFF-Product
— *Kadeřník jí učesal z vlasů*.ORIG-Material *drdol*.PAT-Product
Eng. Hairdresser arranged a bun.PAT-Product from her hair.ORIG-Material

Source-Substance diathesis
Slunce.ACT-Source *vyzařuje teplo*.PAT-Substance
Eng. The sun.ACT-Source radiates heat.PAT-Substance
— *Teplo*.ACT-Substance *vyzařuje ze slunce*.DIR-Source
Eng. Heat.ACT-Substance radiates from the sun.DIR-Source

Object-Direction diathesis ('from where', 'through' or 'to where')
Marta vylezla kopec.PAT-Object
Eng. Martha climbed the mountain.PAT-Object
— *Marta vylezla na kopec*.DIR-Direction
Eng. Martha climbed up the mountain.DIR-Direction

Direction-Location diathesis
Matka umístila dítě do jeslí.DIR-Direction
Eng. Mother put her child into a nursery school.DIR-Direction
— *Matka umístila dítě v jeslích*.LOC-Location
Eng. Mother put her child into a nursery school.LOC-Location

Agent-Location diathesis

Včely.ACT-Agent *se rojí na zahradě*.LOC-Location
Eng. Bees.ACT-Agent are swarming in the garden.LOC-Location
— *Zahrada*.ACT-Location *se rojí včelami*.MEANS-Agent
Eng. The garden.ACT-Location is swarming with bees.MEANS-Agent

Conclusion

For lexicographic description of verbal valency, it is necessary to specify (i) valency frame of each lexical unit, (ii) information on the applicability of a particular set of rules describing the possible diatheses, and (iii) precise formulations of rules. Information (i) and (ii) are stored in the data component whereas (iii) is stored in the grammar component of the valency lexicon.

We distinguish two types of changes in valency structure, which are referred to as g-diatheses and s-diatheses. G-diatheses are prototypically characterized by morphologically marked form of verb in the marked construction, while the mapping between semantic participants of a generalized situation and valency slots remains unchanged, their number and type are identical (the changes in valency frames are limited to morphemic realizations of individual valency slots). On the other hand, s-diatheses are characterized by changes in number and types of valency slots. They are typically limited to verbs of certain semantic classes.

Distinguishing between g-diatheses and s-diatheses in the valency lexicon VALLEX is motivated by the needs of lexicographic work. In case of g-diatheses, the changes in valency frames are regular enough to be treated in the form of general rules (in the grammar component) and as a single verbal lexical unit (for both syntactic constructions) marked with the possibility of a particular type of diathesis. For s-diatheses, separate lexical units are established and interlinked with general rules identifying a relevant type of s-diathesis. This solution reflects the corpus evidence that changes in valency structure of verbs are diverse even within an individual type of s-diathesis.

References

[1] Apresjan, J. D. (1974). *Leksicheskaja semantika. Sinonimicheskie sredstva jazyka*. Nauka, Moskva.
[2] Borer, H. (2005). *The Normal Course of Events*. Oxford University Press, Oxford.
[3] Cholodovič, A. A. (1970). Zalog. Kategoria zaloga. In *Materialy konferencii*, pages 2–26, Leningrad.
[4] Chomsky, N. A. (1957). *Syntactic Structures*. Mouton, The Hague.
[5] Chomsky, N. A. (1965). *Aspects of the Theory of Syntax*. MIT Press, Cambridge.
[6] Chrakovskij, V. S., editor (1977). *Problemy lingvisticheskoj tipologii i struktury jazyka*. Nauka, Leningrad.
[7] Daneš, F. (1968). Some Thoughts on the Semantic Structure of the Sentence. *Lingua*, 21:55–69.
[8] Daneš, F. (1985). *Věta a text: studie ze syntaxe současné češtiny*. Academia, Praha.

[9] Daneš, F., Grepl, M., and Hlavsa, Z., editors (1987). *Mluvnice češtiny 3*. Academia, Praha.
[10] Goldberg, A. E. (1995). *Constructions. A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago.
[11] Grepl, M. and Karlík, P. (1998). *Skladba češtiny*. Votobia, Olomouc.
[12] Levin, B. C. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago Press, Chicago and London.
[13] Lopatková, M., Žabokrtský, Z., and Kettnerová, V. (2008). *Valenční slovník českých sloves*. Nakladatelství Karolinum, Praha.
[14] Mikulová, M. et al. (2006). Annotation on the Tectogrammatical Level in the Prague Dependency Treebank. Annotation Manual. Technical Report TR-2006-30, ÚFAL MFF UK, Prague.
[15] Ondrejovič, S. (1989). *Medzi slovesom a vetou*. Jazykovedné štúdie. Veda, Bratislava.
[16] Panevová, J. (1980). *Formy a funkce ve stavbě české věty*. Academia, Praha.
[17] Panevová, J. (2007). Znovu o reciprocitě. *Slovo a slovesnost*, 68:91–100.
[18] Panevová, J. and Řezníčková, V. (2001). K možnému pojetí všeobecnosti aktantu. In Hladká, Z. and Karlík, P., editors, *Čeština – univerzália a specifika 3*, pages 139–146. Masarykova Universita, Brno.
[19] Panevová, J. et. al (manuscript). *Syntax současné češtiny (na základě anotovaného korpusu)*. Nakladatelství Karolinum, Praha.
[20] Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Reidel, Dordrecht.
[21] Štícha, F. (1984). *Utváření a hierarchizace struktury větného znaku*. Univerzita Karlova, Praha.
[22] Uspenskij, V. A. (1977). K ponjatiju diatezy. In Chrakovskij, V. S., editor, *Problemy lingvisticheskoj tipologii i struktury jazyka*, pages 65–84. Leningrad.

Chapter C

Current Concept of the Valency Lexicon: *VALLEX, Version 2*

Struktura slovníku *VALLEX* [The Structure of the *VALLEX* lexicon]

Introduction to the book *Valenční slovník českých sloves [Valency Lexicon of Czech Verbs]*, Prague, Karolinum, 2008
(co-authors of the Introduction Z. Žabokrtský and V. Kettnerová)

Struktura slovníku *VALLEX*

Obsah *Valenčního slovníku českých sloves VALLEX* zde popíšeme pouze z hlediska jeho struktury. Lingvistické aspekty vyžadující širší vysvětlení či diskusi jsou většinou ponechány stranou, odkazujeme pouze na základní literaturu týkající se dané problematiky.

Odborná terminologie, kterou v textu užíváme, buď patří k ustálené lingvistické terminologii, nebo vychází z terminologie Funkčního generativního popisu (FGP), který slouží jako podkladová teorie *VALLEXu*. Pokud tomu tak není, jsou potřebné termíny zavedeny v textu.

1 Lexémy, lexikální formy a lexikální jednotky

Slovník *VALLEX* je na nejvyšší úrovni tvořen lexémy, kterým odpovídají jednotlivá slovníková hesla. *Lexémem* přitom rozumíme abstraktní jednotku, která v sobě spojuje formální i významovou složku (viz Cruse, 1986), u polysémních/polysémických/víceznačných lexémů též označovanou jako hyperlexém (viz i Filipec – Čermák, 1985, Karlík et al., 2002). Lexém sdružuje množinu všech možných manifestací slovesa v textu/řeči, kterou zde budeme označovat jako množinu všech možných *lexikálních forem*, a množinu *lexikálních jednotek* (LU jako ‚lexical unit'), které reprezentují jeho významové složky, v české terminologii obvykle označované jako jednotlivé lexie či základní lexikální jednotky, viz terminologickou poznámku níže. Lexikální formy jsou tedy všechny morfematické formy slovesa (celé jeho paradigma), viz odd. 2, zatímco lexikální jednotky zhruba odpovídají lexému v jednom určitém významu a nesou informaci o syntaktických a syntakticko-sémantických rysech slovesa v tomto významu, viz odd. 3.

Terminologická poznámka: Zde se odchylujeme od (ne zcela jednotné) české terminologie, kde se obvykle termíny lexém a lexikální jednotka užívají víceméně synonymně, viz Filipec – Čermák (1985), str. 28: „Tento termín [lexém] je tedy synonymní s termínem LJ [= lexikální jednotka], ale je ještě dále diferencován." Každému významu (polysémického) lexému odpovídá jedna základní lexikální jednotka (tamtéž): „Lexikální jednotka jako polysémický lexém je ... útvar zahrnující tolik různých monosémických základních lexikálních jednotek, lexií (...), kolik má různých významů."

Ve *VALLEX*u se přikláníme k terminologii, která je běžná v anglicky psané odborné literatuře, kde je termínem ‚lexical unit' označován koncept odpovídající základní lexikální jednotce (viz Cruse, 1986, Ruppenhofer et al., 2006).

2 Lexikální formy a lemmata

Množiny všech možných lexikálních forem tvořících formální složku jednotlivých lexémů jsou reprezentovány infinitivními tvary slovesa obvykle nazývanými *lemmata*.

Slovníkové heslo je ve *VALLEXu* uvedeno lemmatem, případně seznamem lemmat vztahujících se k danému lexému (včetně příp. morfému *se/si*, viz odd. 2.1). Jednotlivá lemmata jsou doplněna o další informace:

- informace o vidu v horním indexu, viz odd. 2.2;
- římská číslice v dolním indexu rozlišující homografy (viz odd. 2.4).

Slovník *VALLEX* (v návaznosti na teorii FGP) zachycuje valenční chování vidových protějšků v rámci jediného lexému. Proto je slovníkové heslo typicky uvedeno dvěma (příp. více) slovesnými lemmaty, nedokonavým a dokonavým infinitivním tvarem slovesa, viz odd. 2.2. Dalším důvodem, proč se v záhlaví slovníkového hesla může vyskytnout více lemmat, jsou pravopisné varianty slovesa, viz odd. 2.3.

2.1 Reflexivní lemmata

Z hlediska zachycení reflexivity ve slovníku jsou ve *VALLEXu* rozlišovány dva základní typy reflexivních konstrukcí (viz též Karlík et al., 2002):

- **Reflexivní lexémy.** Jako reflexivní lexémy jsou označována inherentně reflexivní slovesa, *reflexiva tantum*, u nichž je morfém *se/si* považován za součást lemmatu.
  Řadí se k nim:
  - primární reflexiva tantum (v Karlík et al., 2002 označovaná jako inherentně reflexivní slovesa), tedy slovesa, která se v nereflexivní formě vůbec nevyskytují (např. *bát se*, *smát se*), či slovesa, která nereflexivní formy sice mají, lexikální jednotky odpovídající reflexivním a nereflexivním formám si jsou však natolik významově vzdáleny, že jsou obvykle vyčleňovány do dvou lexémů (např. *chovat se* vs. *chovat*);
  - tzv. odvozená/sekundární reflexiva (v Karlík et al., 2002 též inherentně reflexivní varianta slovesa), tedy slovesa, kde *se/si* je slovotvorně motivovaným morfémem, reflexivní forma je tudíž v nějakém významovém vztahu k nereflexivní formě, např. vyjadřuje samovolnou či bezděčnou činnost (např. *šířit se*, *vrátit se*).

  Reflexivní lexémy jsou ve *VALLEXu* zachyceny v samostatných slovníkových heslech.
- **Reflexivní užití nereflexivních lexémů.** Pokud reflexivní morfém *se/si* nese syntaktickou funkci, jsou reflexivní formy sloves zachyceny v rámci nereflexivního lexému, kde je též specifikována jejich syntaktická funkce (viz oddíly 5.2 a 5.3):
  - *se* může být součástí tvaru tzv. reflexivního pasiva (např. *pátrá se po zloději*);
  - *se/si* může označovat doplnění obsazující valenční pozici řídícího slovesa u tzv. vlastního reflexiva (kde *se/si* lze nahradit silnou podobou zájmena *sebe/sobě*, např. *vidět se* (= sebe), *darovat si* (= sobě) *dort*, kde *se/si* je PAT (patient), resp. ADDR (adresát) koreferující s ACT (aktorem) řídícího slovesa);
  - *se/si* může mít reciproční funkci (např. *kopat se* v *kopou se vzájemně do nohou*).

Poznámka: Některé lexikální jednotky mají reflexivní i nereflexivní podobu lemmatu beze změny významu, např. *myslím (si), že to tak není* (někdy se označuje se jako volné *se/si*). Tento typ *se/si* bývá zaznamenán pouze jako jeden z příkladů u nereflexivní lexikální jednotky.

2.2 Vid a vidové protějšky

V češtině se pro kategorii vidu rozlišují dvě základní hodnoty, nedokonavost a dokonavost. Vedle toho se vyčleňují též iterativa jako specifická podtřída nedokonavých sloves a slovesa obouvidová (slovesa, která se v určitých kontextech chovají jako dokonavá, v jiných kontextech jako nedokonavá).

Ve *VALLEXu* jsou do jediného lexému spojeny tzv. *pravé vidové dvojice* tvořené sufixálně – nedokonavé sloveso je utvořeno od dokonavé formy slovesa např. příponou *-va(t)* (např. *ochutnat* → *ochutnávat*), příponou *-ova(t)* (např. *dokončit* → *dokončovat*), příponou *-a(t)* (např. *vyrůst* → *vyrůstat*) či příponou *-ě/e(t)* (např. *otočit* → *otáčet*). Dále jsou v jednom lexému zachyceny i supletivní páry (např. *vzít – brát*, *najít – nacházet*).[6] Pokud existuje i běžně užívaná iterativní podoba, je rovněž zahrnuta do příslušného lexému (např. slovesa *nasedat*^{impf}, *nasednout*^{pf} a *nasedávat*^{iter} jsou popsány v jednom slovníkovém hesle odpovídajícím jednomu lexému).

[6] Dokonavé protějšky nedokonavých sloves tvořené prefixálně v rámci stejného lexému zachyceny nejsou. K tomuto rozhodnutí vedly praktické důvody, neboť ne vždy je zcela jednoznačné, které z více možných prefigovaných lemmat považovat za vidový protějšek.

Ve *VALLEXu* se informace o vidu zachycuje u každého lemmatu jako horní index, který může nabývat následujících hodnot:

- *impf* pro nedokonavá slovesa;
- *pf* pro dokonavá slovesa;
- *iter* pro iterativa (násobená slovesa);
- *biasp* pro obouvidová slovesa.

V rámci jediného slovníkového hesla je v některých případech zachyceno více podob jednoho ze členů vidové dvojice (aniž by šlo o varianty, viz odd. 2.3), a to v případech, kdy jednomu lemmatu jedné vidové hodnoty odpovídají různá lemmata s druhou hodnotou (např. dokonavá slovesa *dohonit*^{pf} i *dohnat*^{pf} mají nedokonavý protějšek *dohánět*^{impf}, naopak dokonavé sloveso *odvinout*^{pf} má dva nedokonavé protějšky *odvinovat*^{impf} a *odvíjet*^{impf} – každá tato trojice lemmat reprezentuje jediný lexém, odpovídá jí tedy jediné slovníkové heslo). Při stanovování vidových protějšků se *VALLEX* přidržuje vztahů stanovených ve slovníku SSJČ.

Typicky taková lemmata sdílejí (alespoň jednu) lexikální jednotku, viz odd. 3, i když u nich může docházet k modifikaci významu (spočívající zejména ve změně ‚způsobu slovesného děje', např. u některých sloves pohybu, u sloves distributivních či u sloves momentálních s příponou *-nou(t)*). Např. lemmata *odříznout*^{pf} a *odřezat*^{pf} jsou zahrnuta spolu s lemmatem *odřezávat*^{impf} do jediného lexému.

Poznámka k notaci: V seznamu lemmat reprezentujících lexém se tedy může vyskytnout více lemmat se stejnou vidovou charakteristikou. V takovém případě je tato charakteristika doplněna arabskou číslicí tak, aby index s vidovou charakteristikou mohl sloužit jako jednoznačný identifikátor lemmatu ve slovníkovém hesle. Tento identifikátor uvádí glosy a příklady, případně další údaje, které se vztahují pouze k některým z lemmat reprezentujících lexém. Ve *VALLEXu* se to týká např. sloves *dohánět*^{impf}, *dohnat*^{pf1} *dohonit*^{pf2}; *odvinovat*^{impf1}, *odvíjet*^{impf2}, *odvinout*^{pf}; *odřezávat*^{impf}, *odříznout*^{pf1}, *odřezat*^{pf2}.

Lexikální jednotka se typicky vztahuje ke všem lemmatům daného lexému, která jsou uvedena v záhlaví slovníkového hesla. Toto obecné pravidlo má však řadu výjimek, konkrétní lexikální jednotky (viz odd. 3) se mohou vztahovat jen k některým z uvedených lemmat. Například sloveso *odpovědět*^{pf} je dokonavým protějškem slovesa *odpovídat*^{impf} ve smyslu ‚dávat odpověď', ale již ne ve smyslu ‚reagovat', ‚mít odpovědnost' či ‚být ve shodě/v souladu; korespondovat' (v těchto významech jde o imperfektum tantum). V takovém případě jsou za arabskou číslicí uvádějící příslušnou lexikální jednotku uvedena za značkou jen všechna lemmata, ke kterým se tato lexikální jednotka vztahuje (s vyznačeným videm a případnými variantami a indexem pro homografy, viz níže); toto omezení se neuvádí pro iterativa.

Ve třídě nedokonavých sloves se dále vyčleňuje skupina sloves násobených (iterativní slovesa, iterativa) označujících opakovaný děj. V češtině je tvoření iterativ velmi produktivní a do značné míry pravidelné, tvoří se od nedokonavých sloves příponou *-va(t)* s kvantitativní či kvalitativní změnou vokálu před příponou (např. *volat→ volávat*, *křičet→ křičívat*, *být→ bývat*), příp. příponou *-a(t)* (např. *jíst→ jídat*). Vzhledem k vysoké produktivitě při tvorbě iterativ nejsou ve *VALLEXu* iterativa zachycena vyčerpávajícím způsobem: iterativa s rozšířenou variantou přípony *-váva(t)* (např. *chodívávat*) nejsou uváděna vůbec, z ostatních iterativ jsou uváděna pouze ta, která se ve sledovaných textech a zdrojových slovnících vyskytovala pravidelněji.

Lemmata iterativních sloves jsou součástí záhlaví slovníkového hesla, neuvádějí se však pro ně glosy a příklady ani se nezaznamenává omezení u lexikálních jednotek, pro které iterativum nelze užít.

2.3 Varianty lemmatu

Varianty lemmatu (často ortografické alternativy) jsou chápány jako skupina dvou nebo více lemmat reprezentujících daný lexém, která jsou zaměnitelná v jakémkoliv kontextu beze změny významu, vztahují se k nim tedy stejné lexikální jednotky (např. *dozvědět/dovědět se*, *dýchnout/dechnout*). Obvykle se varianty

liší jen alternací v morfematickém kmenu slovesa, která je případně doprovázena stylistickým posunem (např. *myslet/myslit*), navíc někdy mohou mít obě lemmata společné některé tvary paradigmatu (např. *mysli* (imperativ) je společným tvarem pro *myslet* i *myslit*).

Všechny varianty lemmatu jsou popsány společně v jednom lexému. Varianty jsou uvedeny v záhlaví hesla, jsou oddělené lomítkem, např. *dovědět/dozvědět se*, *dýchnout/dechnout*, *myslet/myslit*.

Přes toto základní vymezení variant *VALLEX* obsahuje řídké výjimky, kdy lze užít pouze jedno z lemmat. Např. lemmata *plavat* a *plovat* jsou tradičně považována za varianty (viz SSJČ), přestože v některých kontextech lze užít jen *plavat*, např. *plavat při zkoušce* vs. **plovat při zkoušce*. V případě, že lze v daném významu užít pouze jednu z variant, jsou za arabskou číslicí uvádějící příslušnou lexikální jednotku za značkou jen uvedena ta lemmata a jejich varianty, ke kterým se tato lexikální jednotka vztahuje (s vyznačeným videm a indexem pro homografy, viz níže).

2.4 Homografy (homonyma)

Jako *homografy* jsou ve *VALLEXu* označována lemmata s identickou grafickou podobou, ale bez zřejmého sémantického vztahu. Jde tedy o různé lexémy, jejichž lemmata jsou reprezentována stejnou kombinací morfémů, tj. mají stejné grafématické vyjádření.

Často se též liší jejich etymologie (např. *nakupovat*$_I$ jako *nakupovat dětem oblečení* vs. *nakupovat*$_{II}$ jako *nakupovat kolem sebe hromady věcí*), vid (např. *stačit*$_I^{impf}$ jako *stačí mu to ke štěstí* či *Petr stačí Pavlovi v běhu* vs. *stačit*$_{II}^{pf}$ jako *stačí dorazit do školy včas*) nebo některé tvary paradigmatu (např. infinitiv *žít* je homograf, který má pro 3. osobu singuláru minulého času formu *žil* pro význam ‚být naživu; trávit čas', např. *Jan žil v Praze*, ale formu *žal* pro význam ‚kosit; sekat', např. *Jan žal trávu*).

Ve *VALLEXu* jsou homografy rozlišovány římskou číslicí v dolním indexu (v případě reflexivních lemmat před morfémem *se/si*, např. *dít*$_I$, *dít*$_{II}$ *se*).

Terminologická poznámka: Zde se držíme terminologie běžné v anglicky psané literatuře, která rozlišuje homografy jako jednotky se stejnou psanou podobou (bez ohledu na podobu zvukovou) a homofony jako jednotky se zvukově stejnou podobou. Termín *homograf* užívaný ve *VALLEXu* tedy zahrnuje termíny homonymum (jednotka se stejnou psanou i zvukovou podobou) i homograf (jednotka se stejnou psanou, ale odlišnou zvukovou podobou), jak je užívá česky psaná literatura.

3 Lexikální jednotky

V koncepci slovníku *VALLEX* reprezentují lexikální jednotky významovou složku, která spolu se složkou formální (s možnými lexikálními formami) vytváří lexém. Každý lexém je tedy tvořen množinou lexikálních jednotek (LU), kterým jsou přiřazeny příslušné lexikální formy (reprezentované lemmaty). V souladu s D. A. Crusem (viz Cruse, 1986) považujeme lexikální jednotky za „komplexní jednotky s (relativně) stálými, diskrétními sémantickými vlastnostmi.“[7] Stručně řečeno, jde zhruba o ‚dané slovo v daném významu'.

Poznámka k vyčleňování jednotlivých LU: Pro vyčleňování jednotlivých významů daného lexému neexistují všeobecně přijatá testovatelná kritéria, přechod od jednoho významu k druhému je v řadě případů pozvolný. Ve *VALLEXu* je při rozlišování jednotlivých LU kladen důraz na syntaktická kritéria, zejména na podobu valenčního rámce, včetně povrchové realizace jednotlivých valenčních doplnění (viz odd. 4). Přitom se ovšem přihlíží též k sémantice.

- Změny ve valenčním rámci (s výjimkou morfematických variant) vedou k vyčlenění více LU, i když je význam těchto LU blízký (např. následující užití slovesa *poslat* bude popsáno dvěma LU: *poslat peníze do banky/na účet.*DIR3 vs. *poslat peníze dětem.*ADDR).

[7] Cruse (1986): “form-meaning complexes with (relatively) stable and discrete semantic properties”.

- Podobně vede k vyčlenění různých LU i různá syntaktická strukturace (např. *naložit vůz.PAT senem.EFF* vs. *naložit seno.PAT na vůz.DIR3*; *žnout louku.PAT* vs. *žnout trávu.PAT na louce.LOC*). Systematické provázání těchto blízkých LU pomocí tzv. alternačního modelu navrženého ve studiích Žabokrtský (2005) a Lopatková et al. (2006) není v současné verzi slovníku uplatněno.
- Jednotlivé valenční členy jsou specifikovány syntakticko-sémantickým vztahem k řídícímu slovesu, odlišnost tohoto vztahu opět vede k různým LU (např. tři LU pro lexém *pocházet*: *rukopis pochází ze 14. stol.TFRWH* vs. *pochází z venkova.DIR1* vs. *všechno zlo pochází z bídy a neznalosti.PAT*).
- Pokud má sloveso dva (či více) zřetelně odlišné významy, jsou tyto významy popsány různými LU i v případech, kdy se valenční rámec neliší (např. dvě LU pro sloveso *chovat*: *chovat dítě.PAT v náručí.LOC* vs. *chovat prasata.PAT na farmě.LOC*).

Terminologická poznámka: Lexikální jednotka spolu se svými lexikálními formami, jak je chápána ve *VALLEXu*, odpovídá jednotce v české tradici označované jako monosémní/monosémický lexém, lexie nebo též základní lexikální jednotka (viz Filipec – Čermák, 1985, Karlík et al., 2002), zde též odd. 1.

Lexikální jednotky jsou ve *VALLEXu* číslovány arabskými číslicemi – pokud má lexém více významů popsaných několika lexikálními jednotkami, je každá lexikální jednotka uvozena touto číslicí.

Pořadí lexikálních jednotek není zcela arbitrární, není ale přísně systematické. V této podobě slovníku je dáno intuicí autorů (s přihlédnutím ke vzorku korpusového materiálu) – primární a/nebo velmi frekventované významy jsou uváděny na prvních místech, zatímco řídké a idiomatické významy jsou řazeny na konec slovníkového hesla.

Pokud není specifikováno jinak, vztahuje se lexikální jednotka ke všem lemmatům reprezentujícím daný lexém (uvedeným v záhlaví slovníkového hesla). V případech, kdy se daná lexikální jednotka vztahuje jen k některým lemmatům ze seznamu v záhlaví, jsou za značkou jen uvedena všechna lemmata, ke kterým se tato lexikální jednotka vztahuje (s vyznačeným videm a případnými variantami a indexem pro homografy).[8]

Lexikální jednotky zhruba odpovídají lexému v určitém významu a nesou informaci o syntaktických a sémantických rysech slovesa v daném významu. Příslušné informace jsou ve *VALLEXu* zachyceny jako povinné a nepovinné atributy lexikální jednotky. Povinné atributy musí být vyplněny pro každou lexikální jednotku. Nepovinné atributy mohou být nevyplněny, buď protože se u dané lexikální jednotky nevyskytují (např. kontrola se uvádí jen u sloves s touto vlastností, viz odd. 5.1), nebo protože dané informace nejsou v současné podobě dostupné (např. určení syntakticko-sémantické třídy slovesa, viz odd. 5.4).

Povinné atributy lexikální jednotky:

- valenční rámec, viz odd. 4;
- glosa – sloveso nebo parafráze charakterizující daný význam slovesa; glosy nelze pokládat za synonyma nebo dokonce za lexikografické definice, slouží pouze pro orientaci ve slovníkovém hesle;
- příklad – věty nebo fragmenty vět obsahujících dané sloveso v daném významu, případně s označením zdroje příkladu, např. ČNK, SSJČ apod.

Nepovinné atributy lexikální jednotky:

- kontrola, viz odd. 5.1;
- možný typ reflexivních konstrukcí, viz odd. 5.2;
- možný typ recipročních konstrukcí, viz odd. 5.3;
- příslušnost k syntakticko-sémantické třídě, viz odd. 5.4;
- označení idiomu, viz odd. 5.5.

[8] Pokud se daná lexikální jednotka vztahuje ke všem lemmatům uvedeným v záhlaví slovníkového hesla s výjimkou iterativa, žádné omezení se neuvádí.

4 Valenční rámce

Nejdůležitější sémanticko-syntaktická charakteristika slovesa je zachycena ve formě *valenčního rámce*. Valenční rámec (v užším smyslu) ve FGP sestává z aktantů (obligatorních i fakultativních) a z obligatorních volných doplnění, v novějších studiích je pak obohacen o tzv. kvazivalenční doplnění. Ve *VALLEXu* se kromě členů takto pojímaného valenčního rámce uvádí i nevelké množství fakultativních volných doplnění (dále typická doplnění). S daným slovesem se mohou vyskytovat též ostatní volná doplnění, ta však nejsou ve valenčním rámci uváděna, neboť jejich výskyt není podle FGP podmíněn syntakticky. Klasifikaci valenčních doplnění tvořících obohacený valenční rámec ve *VALLEXu* je zde věnován oddíl 4.1.

Ve *VALLEXu* jsou valenční rámce modelovány jako posloupnosti valenčních a nevalenčních pozic, kde každá pozice odpovídá jednomu valenčnímu, příp. typickému doplnění daného slovesa. Každá pozice je charakterizována:

- funktorem, viz odd. 4.1;
- seznamem možných morfematických forem, viz odd. 4.2;
- informací o obligatornosti, viz odd. 4.3.

Jistá volná doplnění se systematicky objevují společně. Tato pravidelnost je zachycena pomocí mechanismu expanze valenční či nevalenční pozice, viz odd. 4.4; plný valenční rámec se získá expanzí pozice uváděné ve slovníku.

4.1 Aktanty a volná doplnění

Ve valenční teorii FGP se slovesná doplnění dělí na *aktanty* (vnitřní doplnění ve všech svých výskytech) a *volná doplnění* (viz zejména Panevová, 1974, 1980, 1994). Zkratky pro jednotlivé aktanty a volná doplnění se dále souhrnně označují jako *funktory*. Jednotlivé funktory ve *VALLEXu* tedy označují typ sémanticko-syntaktického vztahu mezi slovesem a jeho doplněním.

Aktanty jsou určovány převážně na základě syntaktických pravidel:

- počet pozic pro aktanty je charakteristický pro každé sloveso a pro každé sloveso tedy musí být vymezen ve slovníku;
- jako rozvití nějakého konkrétního slovesa se daný aktant vyskytuje nejvýše jednou (vyjma případů souřadnosti a apozice).

Doplňme, že aktanty jsou doplnění typicky rekční.

Empiricky bylo stanoveno pět aktantů: aktor (ACT), patient (PAT), výsledek děje (EFF), adresát (ADDR) a původ (ORIG). Zásady pro určování jednotlivých aktantů lze nalézt např. v Panevová – Skoumalová (1992), nově v Mikulová et al. (2005), kde je i jejich (zatím nejpodrobnější) charakteristika; zde je krátká charakteristika aktantů uvedena níže.

Volná doplnění jsou na rozdíl od aktantů sémanticky distinktivní. Charakterizují je následující vlastnosti:

- omezení na slučitelnost slovesa s volnými doplněními nemají podle FGP z velké části syntaktický charakter;
- dané sloveso může být rozvíjeno jedním typem volného doplnění i více než jedenkrát.

Volná doplnění (včetně typických forem) jsou popsána v Mikulová et al. (2005), níže uvádíme typické příklady.

Dichotomie aktant – volné doplnění byla v novějších studiích obohacena o třetí typ tzv. *kvazivalenčních doplnění* (viz Panevová, 2003, Lopatková – Panevová, 2006), která jsou na hranici mezi aktanty

a volnými doplněními. Jde o doplnění rozvíjející relativně uzavřenou (sémanticky homogenní) třídu sloves, jsou to doplnění rekční a dané doplnění nelze u jednoho řídícího slovesa opakovat. Podobně jako volná doplnění jsou však sémanticky distinktivní a typicky se nejedná o doplnění obligatorní. Jde např. o záměr INTT u sloves pohybu (třídy motion a transport, odd. 5.4, např. *Petr jel nakoupit*, *Maruška šla na jahody*) či o překážku OBST (třída contact, např. *zakopl o kořen, zachytil šálou o hřebík*).

Dále se ve *VALLEXu* uvádí nevelké množství fakultativních volných doplnění, která obvykle nespecifikují význam slovesa, ale typicky se vztahují k celé syntakticko-sémantické třídě sloves. Pro některá doplnění mají prototypickou formu (např. instrumentál pro způsob, *psal perem, jel vlakem*, či předložková skupina pro+4 pro benefaktiv, *dělal to pro děti*), jindy je jejich forma dána sémantikou příslušného doplnění (např. směrová doplnění DIR1, DIR2 a DIR3 u sloves pohybu, více viz odd. 4.2). Takováto doplnění se obvykle chápou jako doplnění nevalenční, ve *VALLEXu* se však uvádějí, protože tato informace může být s úspěchem využita při automatické analýze češtiny.

Terminologická poznámka: Aktanty a obligatorní volná doplnění ve FGP víceméně odpovídají konstitutivním větným členům (obligatorním i potenciálním), typická a kvazivalenční doplnění odpovídají větným členům nekonstitutivním (viz Daneš, 1971, Daneš – Hlavsa, 1987, Grepl – Karlík, 1998). V termínech Mluvnice češtiny 3 (Daneš et al., 1987) odpovídají aktanty a obligatorní volná doplnění participantům intenčního pole.

Dělení na aktanty a volná doplnění se také víceméně shoduje s dělením na argumenty a adjunkty podle Grepl – Karlík (1998), Karlík et al. (2002).

Funktory označující typ sémanticko-syntaktického vztahu jsou blízké tzv. hloubkovým pádům/rolím C. J. Fillmora (např. Fillmore, 1969) či theta rolím podle N. Chomského (viz např. Chomsky, 1981, Jackendoff, 1990).

Charakteristika aktantů

- **Aktor/konatel (funktor ACT).** Valenční doplnění aktor je (levovalenční) aktant, který je vymezen jako první aktant slovesa – označuje doplnění zaplňující první syntaktickou pozici slovesa (např. *maminka.ACT upekla koláč*, *voda.ACT naplnila jámu*, *kniha.ACT vyšla*). V zásadě je to doplnění v pozici syntaktického subjektu u aktivní konstrukce (v případě pasivní konstrukce se jedná o doplnění se stejným sémantickým vztahem ke slovesu, např. *nakladatelství.ACT Odeon vydalo knihu* i *kniha byla vydána nakladatelstvím.ACT Odeon*). Jde o rozšířené pojetí konatele děje zahrnující jak činitele, tak i nositele stavu/děje a příbuzné sémantické role.
  Je-li jeden z aktantů vyjádřen dativní formou (a druhý nominativní formou), přihlíží se též k sémantice aktantu. Vyjadřuje-li aktant s dativní formou proživatele, hodnotí se tento aktant jako aktor (a aktant v nominativu jako patient) (např. *kniha se mi.ACT líbila*).
- **Patient (funktor PAT).** Valenční doplnění patient je (pravovalenční) aktant, který je vymezen jako druhý aktant slovesa – označuje doplnění zaplňující druhou syntaktickou pozici slovesa (např. *Marie postavila vázu.PAT na stůl*, *maminka upekla koláč.PAT*, *kniha patří Janovi.PAT*, *obraz.PAT se mi nelíbil*, *vzdal se odměny.PAT*, *učil se zahradníkem.PAT*, *vyprávěl nám o dovolené.PAT*). V zásadě je to doplnění v pozici přímého (syntaktického) objektu u aktivní konstrukce (v případě pasivní konstrukce se jedná o doplnění se stejným sémantickým vztahem, např. *nakladatelství Odeon vydalo knihu.PAT*, *kniha.PAT byla vydána nakladatelstvím Odeon*). Jde o rozšířené pojetí předmětu zasaženého dějem.
- **Výsledek děje, efekt (funktor EFF).** Valenční doplnění efekt je (pravovalenční) aktant, který se uplatňuje u sloves se třemi (a více) syntaktickými pozicemi. Funktor EFF je přiřazován zejména doplněním obsazujícím třetí syntaktickou pozici u sloves, která odpovídá jednak tzv. doplňku u sloves neplnovýznamových (viz též Šmilauer, 1966),[9] jednak jde o druhý věcný předmět sémanticky se blížící výsledku děje. Obecně vyjadřuje vlastnost nebo stav, které má doplnění s funktorem PAT za

[9] Doplněk doplňovací v prvním vydání Novočeské skladby (Šmilauer, 1947).

jistého děje nebo které se mu jistým dějem přisuzují (např. *považoval Pavla za odborníka.EFF, jmenovali ho ředitelem.EFF, my tomu říkáme efekt.EFF sněhové koule, Petr přeložil knihu do češtiny.EFF, svazovali kmeny do voru.EFF*).

- **Adresát (funktor ADDR).** Valenční doplnění adresát je (pravovalenční) aktant, který je vymezen jako aktant slovesa typicky vyjadřující roli příjemce děje (např. *dal dceři.ADDR k narozeninám knížku, řekl synovi.ADDR pravdu, bratrovi.ADDR nezaplatili dohodnutou mzdu, celé dětství soupeřil o matčinu přízeň s bratrem.ADDR*). Funktor ADDR se uplatňuje u sloves se třemi (a více) syntaktickými pozicemi. Jeho typickým rysem je životnost. Prototypicky jde o doplnění v pozici nepřímého objektu (např. *předal knihu Janovi.ADDR, kniha byla předána Janovi.ADDR*).
- **Původ (funktor ORIG).** Valenční doplnění původu je (pravovalenční) aktant, který je vymezen jako aktant slovesa vyjadřující roli původu (např. *vyrábějí ze dřeva.ORIG stoly i židle, slyšel o neštěstí od sousedů.ORIG, nevzal od něj.ORIG za práci peníze*). Funktor ORIG se uplatňuje u sloves se třemi (a více) syntaktickými pozicemi.

Poznámka: Podle valenční teorie FGP, viz zejména Panevová (1974, 1980), platí pro určování funktorů následující princip, který je označován jako *princip posouvání* (shifting): pokud má sloveso jediný aktant, jde o aktor, sloveso se dvěma aktanty má vždy aktor a patient; teprve u sloves se třemi a více aktanty přistupují při výběru funktoru sémantická kritéria.

Funktory ve *VALLEXu*. V následujícím výčtu jsou shrnuty funktory, které se v tomto slovníku vyskytují. Pro úplnost zde uvádíme všechny funktory pro volná doplnění s alespoň jedním výskytem ve *VALLEX*u bez ohledu na to, zda jsou v konkrétních příkladových větách příslušná doplnění valenční, nebo zda jde pouze o doplnění typická, tedy nevalenční.

Aktanty:

- ACT (aktor): *Petr čte dopis.*
- PAT (patient): *Potkal jsem bratra.*
- EFF (výsledek děje, efekt): *Jmenovali ho ředitelem.*
- ADDR (adresát): *Petr dal Marii knihu.*
- ORIG (původ): *Upekla z jablek koláč.*

Kvazivalenční doplnění:

- DIFF (rozdíl): *Hodnota akcií stoupla o 100 %.*
- INTT (záměr): *Přišel navštívit Janu.*
- OBST (překážka): *Chlapec zakopl o kořen.*

Volná doplnění (abecedně):

- ACMP (doprovod): *Matka tam šla s dětmi.*
- AIM (účel): *Jan šel do pekárny pro chléb.*
- BEN (benefaktiv): *Připravila snídani pro děti.*
- CAUS (příčina): *Petr pro nemoc končí s prací.*
- COMPL (doplněk): *Pracoval jako učitel.*
- CRIT (kritérium): *Třídili diamanty podle velikosti.*
- DIR1 (směr – odkud): *Přišel z lesa promočený.*
- DIR2 (směr – kudy): *Vydal se do sousední vesnice přes les.*

- DIR3 (směr – kam): *Vydal se do sousední vesnice přes les.*
- DPHR (závislá část frazému): *Novináři ho neustále chytali za slovo.*
- EXT (míra): *Tatínek měřil 2 metry.*
- HER (dědictví): *Pojmenovali nejstaršího syna po otci.*
- LOC (místo): *Narodil se v Itálii.*
- MANN (způsob): *Choval se k ní laskavě.*
- MEANS (prostředek): *Napsal dopis rukou.*
- RCMP (náhrada): *Koupila si nové tričko za 350 Kč.*
- REG (zřetel): *Situace se v tomto ohledu výrazně zlepšila.*
- SUBS (substituce): *Startoval za Slávii.*
- TFHL (čas – na jak dlouho): *Přerušil studium na rok.*
- TFRWH (čas – ze kdy): *Jeho špatné vzpomínky pocházejí právě z tohoto období.*
- THL (čas – jak dlouho): *Strávili jsme tam tři týdny.*
- TOWH (čas – na kdy): *Odložili zkoušku z pondělka na úterý.*
- TSIN (čas – od kdy): *Lhůtu počítáme od okamžiku dodání.*
- TTIL (čas – do kdy): *Potrvá to do večera.*
- TWHEN (čas – kdy): *Babička přijde zítra.*

Poznámka: Kromě těchto funktorů se ve *VALLEXu* vyskytuje ještě hodnota DIR. Ta je však užívána jen v souvislosti s expanzí pozice valenčního rámce, viz odd. 4.4.

Množina funktorů, se kterými se pracuje ve FGP a která je využita v PDT, je bohatší, viz např. Mikulová et al. (2005). Některé z těchto funktorů se však nevyskytují u slovesných doplnění (např. MAT – partitiv, jako ve spojení *sklenice piva.MAT*), jiné funktory specifikují vztahy, které nejsou závislostní (např. koordinaci, *Petr nebo.DISJ Marie*). Další funktory reprezentují závislostní vztahy u sloves, nemají však nikdy valenční povahu (např. ATT – postoj, *udělal to dobrovolně.ATT*).

4.2 Morfematické vyjádření

Každá valenční i nevalenční pozice může být ve větě vyjádřena omezenou množinou výrazových prostředků, morfematických forem. Ve *VALLEXu* je množina možných forem specifikována buď explicitně, nebo implicitně.

U explicitně zachycených forem jsou možné morfematické formy dány výčtem u dané pozice valenčního rámce (dolní index u příslušného funktoru). U aktantů a kvazivalenčních doplnění je tento seznam forem úplný (udávají se formy pro užití slovesa v aktivním tvaru) – jiné prostředky nelze pro vyjádření těchto valenčních doplnění užít.[10] V případě volných doplnění jsou explicitně uvedené formy pro dané sloveso pouze typické, lze užít i další formy dané sémantikou doplnění.

U implicitně zachycených forem se předpokládá, že množina možných forem je dána sémantikou doplnění, tedy vyplývá z příslušného funktoru. Jinými slovy, doplnění se může realizovat jakoukoliv formou vyjadřující daný typ doplnění; její výběr je ovšem závislý na lexikálním obsazení a kontextových podmínkách, např. *bydlí na kopci* vs. *ve vesnici*, *napsal dopis rukou* vs. *na počítači*.

Explicitně zachycené formy. Seznamy morfematických forem, které se mohou vyskytnout u jednotlivých valenčních pozic, sestávají z následujících typů hodnot:

[10] Zcela stranou jsou však ponechány formy, které jsou dány gramatickými pravidly, např. pasivizací nebo reciproční konstrukcí, a dále formy pro partitiv (*dodat sůl – dodat soli*), distributivnost (*rozdal jim jablíčka – rozdal jim po jablíčku*) či méně přesnou kvantifikaci (*přišlo padesát lidí – přišlo na padesát lidí, přišlo okolo padesáti lidí*).

- **Bezpředložkové pády.** Jednotlivé pády jsou označeny příslušnými číslicemi: 1 – nominativ, 2 – genitiv, 3 – dativ, 4 – akuzativ, 5 – vokativ, 7 – instrumentál.
- **Předložkové skupiny.** Jsou určeny lemmatem předložky (v její nevokalizované podobě) a číslem pádu (např. z+2, na+4, o+6, ...). Ve *VALLEXu* se vyskytují následující předložky: *bez, do, jako,*[11] *k, kolem, mezi, místo, na, nad, o, od, po, pod, podle, pro, proti, před, přes, při, s, u, v, z, za*.
- **Infinitivní konstrukce.** Značka inf reprezentuje valenční doplnění ve formě infinitivu slovesa (ve vzácných případech též se spojkou než+inf).
- **Závislé věty.** Závislé věty obsahové uvozené podřadicí spojkou jsou reprezentovány lemmatem této spojky; ve *VALLEXu* se vyskytují následujícíci spojky: aby, ať, až, jak, zda,[12] že.
  Závislé věty obsahové, které nejsou uvozeny spojkami (např. nepřímé otázky uvozené tázacím zájmenem nebo adverbiem), jsou reprezentovány zkratkou cont.
- **Konstrukce s adjektivy.** Zkratka adj-číslice specifikuje doplnění ve formě přídavného jména v příslušném pádu (např. adj-1 pro *cítím se slabý*).
- **Konstrukce s *být*.** Infinitiv slovesa *být* se může vyskytnout v konstrukci s adjektivem či v bezpředložkovém pádu (např. být+adj-1 pro *zdá se to být dostatečné*).
- **Část frazému.** U frazeologických jednotek platí, že pokud je množina lexikálních forem, které naplňují určitou valenční pozici, omezená (často jednočlenná), jsou ve *VALLEXu* uvedeny přímo tyto lexikální formy (např. napospas pro frazém *ponechat napospas*).

Implicitně zachycené formy. Pokud není pro valenční pozici explicitně určena možná forma doplnění, potom množina možných forem vyplývá z funktoru pro toto doplnění. Následující výčet udává formy obvyklé pro dané funktory (seznam vychází z nejčastějších forem pro jednotlivé funktory v PDT).

- ACMP: bez+2, s+7, společně s+7, spolu s+7, v čele s+7, v souvislosti s+7, ve spojení s+7, včetně+2, ...;
- AIM: aby, ať, do+2, k+3, na+4, o+4, pro+4, pro případ+2, proti+3, v zájmu+2, za+4, za+7, že, ...;
- BEN: 3, na+4, na účet+2, na úkor+2, na vrub+2, pro+4, proti+3, v+4, ve prospěch+2, v rozporu s+7, v zájmu+2, ...;
- CAUS: 7, aby, adverb, díky+3, jelikož, ježto, kvůli+3, na+4, na+6, na základě+2, nad+7, následkem+2, od+2, pod+7, pod náporem+2, pod tíhou+2, pod váhou+2, poněvadž, pro+4, protože, v+6, v důsledku+2, v souvislosti s+7, vinou+2, vlivem+2, vzhledem k+3, z+2, z důvodu+2, za+4, za+7, zásluhou+2, že, ...;
- CRIT: 2, 7, dle+2, na+6, na základě+2, po vzoru+2, podle+2, přiměřeně+3, v+6, v duchu+2, v rozporu s+7, v souladu s+7, v souhlase s+7, v závislosti na+6, ve shodě s+7, ve smyslu+2, ve světle+2, z titulu+2, ...;
- DIR1: adverb, od+2, s+2, z+2, ze strany+2, zpod+2, zpoza+2, zpřed+2, ...;
- DIR2: 7, adverb, cestou+2, kolem+2, mezi+7, napříč+7, po+6, podél+2, přes+4, skrz+4, v+6, ...;
- DIR3: 7, adverb, do+2, do čela+2, k+3, kolem+2, mezi+4, mimo+4, na+4, na+6, nad+4, naproti+3, okolo+2, po+4, po+6, pod+4, proti+3, před+4, přes+4, směrem do+2, směrem k+3, směrem na+4, v+4, vedle+2, za+4, za+7, ...;
- EXT: 2, 4, 7, adverb, do+2, k+3, kolem+2, na+4, na+6, nad+4, okolo+2, po+6, pod+7, přes+4, v+4, z+2, za+4, ...;
- LOC: adverb, blízko+2, blízko+3, daleko+2, do+2, kolem+2, mezi+7, mimo+4, na+4, na+6, na úroveň+2, nad+7, naproti+3, nedaleko+2, okolo+2, po+6, poblíž+2, pod+7, podél+2, proti+3, před+7, přes+4, při+6, stranou+2, u+2, uprostřed+2, uvnitř+2, v+6, v čele+2, v oblasti+2, v rámci+2, v řadě+2, vedle+2, za+4, za+7, ...;
- MANN: 7, adverb, do+2, formou+2, na+4, na+6, nad+4, o+4, po+6, pod+7, proti+3, před+7, při+6, přes+4, s+7, v+4, v+6, v podobě+2, ve formě+2, vedle+2, z+2, za+4, za+7, jak, že, ...;
- MEANS: 7, adverb, cestou+2, díky+3, do+2, na+4, na+6, o+6, po+6, pod+7, pomocí+2, prostřednictvím+2, přes+4, s+7, s pomocí+2, skrz+2, v+6, z+2, za+4, za pomoci+2, že, ...;

[11] Slovo *jako* je sice tradičně považováno za spojku, zde je však zahrnuto mezi předložkami, neboť konkrétní valenční doplnění uvozené touto spojkou vyžaduje vždy určitý pád substantiva.

[12] Spojka zda reprezentuje též spojku jestli.

- REG: 7, adverb, bez ohledu na+4, bez zřetele k+3, k+3, kolem+2, na+4, na+6, na téma+2, nad+7, nezávisle na+6, o+6, ohledně+2, po+6, pro+4, před+7, při+6, s+7, s ohledem na+4, se zřetelem k+3, se zřetelem na+4, u+2, v+6, v otázce+2, v případě+2, v rámci+2, v souvislosti s+7, ve věci+2, ve vztahu k+3, vůči+3, vzhledem k+3, z+2, z hlediska+2, za+4, . . . ;
- SUBS: jménem+2, místo+2, namísto+2, výměnou za+4, za+4, . . . ;
- TFHL: adverb, do+2, na+4, po+2, pro+4, . . . ;
- TFRWH: od+2, z+2, . . . ;
- THL: 2, 4, 7, adverb, až, dokud, do+2, na+4, po+4, po dobu+2, přes+4, v+2, za+4, . . . ;
- TOWH: adverb, do+2, k+3, na+4, pro+4, . . . ;
- TSIN: adverb, od+2, počínaje+7, z+2, . . . ;
- TTILL: adverb, do+2, dokud, k+3, než, po+4, . . . ;
- TWHEN: 2, 4, 7, adverb, až, do+2, jakmile, k+3, když, kolem+2, koncem+2, mezi+7, na+4, na+6, na závěr+2, než, o+6, okolo+2, po+6, počátkem+2, postupem+2, poté co, před+7, předtím než, při+6, s+7, u příležitosti+2, v+4, v+6, v době+2, v období+2, v průběhu+2, v závěru+2, z+2, za+2, za+4, začátkem+2, . . .

4.3 Atribut obligatornosti slovesného doplnění

Ve *VALLEXu* se v souladu s valenční teorií FGP slovesná doplnění dělí na obligatorní a fakultativní. *Obligatorností* se rozumí povinná přítomnost daného doplnění v hloubkové (tektogramatické) struktuře, a to bez ohledu na jeho možnou povrchovou vypustitelnost ve větě, viz poznámku níže. Jako kritérium obligatornosti byl stanoven *dialogový test* (viz Panevová, 1974, Sgall et al., 1986). Tento test slouží pro určení obligatornosti doplnění, je-li zkoumaný člen v povrchové větě vypuštěn – např. test obligatornosti doplnění směru-kam (funktor DIR3) pro sloveso *přijít* simuluje dialog mluvčího A a B:

A: *Přátelé už přišli.*
B: *Kam?*
A: **Nevím.*

Odpověď mluvčího A činí dialog deviantní (A musí vědět, o jakém místě mluví) – sloveso *přijít* má tedy obligatorní doplnění DIR3.

Opozice obligatornosti a fakultativnosti se týká aktantů, kvazivalenčních doplnění i volných doplnění.

Poznámka: Některá doplnění obligatorní na rovině významové stavby mohou být vypuštěna (elidována) v povrchové realizaci věty, aniž dojde k porušení gramatičnosti věty (lze říci, že dané doplnění má nulovou lexikální realizaci). K takové elipse dochází tehdy, je-li možné příslušný aktant či volné doplnění snadno doplnit z kontextu, např. *Děti už přišly* (= na místo dané kontextem/sem.DIR3) *a jsou celé promrzlé* (= děti.ACT), případně pokud je daná pozice realizována nějakým typem všeobecného aktantu, např. *Do této buchty se dává sůl, Psali to v novinách* (= všeobecný aktor), viz Daneš (1971), Panevová – Řezníčková (2001).

U každé pozice valenčního rámce je ve *VALLEXu* kódována informace o obligatornosti či fakultativnosti daného doplnění. Obligatorní doplnění (aktanty, kvazivalenční i volná doplnění) jsou tištěna zvýrazněným písmem. Stejným písmem jsou tištěny i fakultativní aktanty a kvazivalenční doplnění (patří též do úzce chápaného valenčního rámce), ty jsou navíc odlišeny značkou opt v horním indexu. Typická volná doplnění, která rozšiřují tradiční valenční rámec, jsou tištěna obyčejným písmem a označena horním indexem typ.

4.4 Expanze valenční pozice

Jistá volná doplnění se systematicky objevují společně. Např. slovesa pohybu lze často rozvíjet všemi typy směrových doplnění, tedy DIR1 (směr-odkud), DIR2 (směr-kudy) a DIR3 (směr-kam). Tato pravidelnost je zachycena pomocí mechanismu expanze pozice valenčního rámce. Pokud je u některé pozice uveden symbol pro expanzi ↑ před funktorem, je plný valenční rámec získán expanzí dané pozice rámce.

Ve *VALLEXu* se symbol pro expanzi ↑ vyskytuje u funktorů DIR, DIR1, DIR2, DIR3 a THL, expanze je popsána následujícími pravidly:

- ↑DIR^{typ} → DIR1^{typ} DIR2^{typ} DIR3^{typ}
  Typické doplnění ↑DIR expanduje ve tři typická doplnění DIR1, DIR2 a DIR3;
  např. rámec pro sloveso *jít* vznikne následující expanzí:
  ACT$_1$ INTT$^{opt}_{k+3,na+4,inf}$ MANN^{typ} MEANS^{typ} ↑DIR^{typ} →
  → ACT$_1$ INTT$^{opt}_{k+3,na+4,inf}$ MANN^{typ} MEANS^{typ} DIR1^{typ} DIR2^{typ} DIR3^{typ}
  (*Petr.ACT jel nakoupit.INTT autem.MEANS z domova.DIR1 přes celou Prahu.DIR2 do Makra.DIR3*)

  Obdobně i pro další typy expanze:
- ↑DIR1 → DIR1 DIR2^{typ} DIR3^{typ}
  Doplnění ↑DIR1 expanduje v obligatorní doplnění DIR1 a typická DIR2 a DIR3.
- ↑DIR2 → DIR2 DIR1^{typ} DIR3^{typ}
  Doplnění ↑DIR2 expanduje v obligatorní doplnění DIR2 a typická DIR1 a DIR3.
- ↑DIR3 → DIR3 DIR1^{typ} DIR2^{typ}
  Doplnění ↑DIR3 expanduje v obligatorní doplnění DIR3 a typická DIR1 a DIR2.
- ↑THL→ TSIN^{typ} THL TTIL^{typ}
  Doplnění ↑THL expanduje ve tři typická doplnění TSIN, THL a TTILL;
  např. rámec pro sloveso *trvat* vznikne následující expanzí:
  ACT$_1$ PAT$^{opt}_3$ ↑THL → ACT$_1$ PAT$^{opt}_3$ THL TSIN^{typ} TTIL^{typ}
  (*Práce na novém obraze.ACT mu.PAT trvala půl roku.THL od jara.TSIN až do konce října.TTILL*)

5 Doplňující syntaktické informace

Jednotlivé LU mohou být obohaceny o nepovinné, doplňující syntaktické, případně syntakticko-sémantické informace, které s valencí souvisejí jen volně. Ve *VALLEXu* je zachycena kontrola (odd. 5.1), reflexivita (odd. 5.2) a reciprocita (odd. 5.3) – jde o gramatické jevy, které přímo ovlivňují povrchové projevy valence. Dále se u vybraných LU uvádí jejich zařazení do syntakticko-sémantické třídy (odd. 5.4), které umožňuje zkoumat, jak se sémantická blízkost sloves odráží v jejich valenčních vlastnostech, a příznak pro idiom (odd. 5.5), neboť frazémy a idiomy často vykazují specifické valenční chování.

5.1 Kontrola

Termínem *kontrola* (značka control) se v tomto kontextu rozumí vlastnost některých sloves (tzv. sloves kontroly) vyžadovat koreferenci mezi svým valenčním doplněním (‚controller‘) a valenčním doplněním podřízeného slovesa (‚controllee‘), viz Panevová (1996). Ve *VALLEXu* je tento vztah zaznamenán pouze pro slovesa, která mohou mít doplnění ve formě infinitivu (bez ohledu na jeho funktor). Za kontrolovaný člen (controllee) je pak považován subjekt tohoto infinitivu (který se v povrchové podobě věty v češtině nevyjadřuje), kontrolující člen (controller) je výraz s ním koreferenční, typicky člen valenčního rámce řídícího slovesa kontroly. Ve *VALLEXu* je kontrola zachycena v atributu control následujícím způsobem:

- koreferenční vztah mezi (nevyjádřeným) subjektem infinitivu a jedním z členů valenčního rámce řídícího slovesa kontroly – atribut control má hodnotu funktoru tohoto valenčního doplnění;
- ostatní případy (tj. pokud takový člen valenčního rámce řídícího slovesa neexistuje) – atribut control má hodnotu ex.

Příklady:

- *pokusit se* (např. *Jiří.ACT se pokusí přijít*) – control: ACT;

- *slyšet* (např. *děti.ACT slyší někoho.PAT přicházet.EFF*) – control: PAT;
- *doporučit* (např. *doporučili mu.ADDR jít.PAT k lékaři*) – control: ADDR;
- *jít* (např. *jde to udělat*, ve smyslu *lze to udělat*) – control: ex.

5.2 Reflexivita

Nepovinný atribut *reflexivity* (značka rfl) udává možnou syntaktickou funkci reflexivního morfému *se/si*, který je v češtině (kromě jiného, viz poznámku níže) formálním prostředkem pro vyjádření následujících syntaktických konstrukcí:

- **sekundární diateze:** částice *se* je součástí reflexivní formy slovesné (viz Karlík et al., 2002), a tedy součástí tvaru tzv. reflexivního pasiva:
  - pro tranzitivní slovesa (slovesa s akuzativní vazbou) (např. *připravovat – plány se připravují*, *bojovat – bojovala se těžká bitva*); atribut rfl má hodnotu pass;
  - pro intranzitivní slovesa (např. *pátrat – pátrá se po zloději*, *chodit – v neděli se chodí do kostela*, *bojovat – s nepřáteli se nakonec nebojovalo*); atribut rfl má hodnotu pass0;
- **gramatická koreference:** zájmeno *se/si* zaujímá pozici valenčního doplnění, které je koreferenční se jménem v subjektu a vyjadřuje, že subjekt vykonává děj sám na sobě; jde o tzv. vlastní reflexiva:
  - je-li příslušná valenční pozice zaplňovaná doplněním s akuzativní formou (a jde tedy o formu zájmena *se*), má atribut rfl hodnotu cor4 (např. *mýt se* (= sebe), *vidět se* (= sebe), *darovat se* (= sebe), kde *se* je patient (PAT) koreferující s aktorem (ACT) řídícího slovesa *mýt*, *vidět* a *darovat*);
  - pro valenční doplnění s dativní formou (a tedy formou zájmena *si*) má atribut rfl hodnotu cor3 (např. *darovat si (dort)* (= sám sobě), kde *si* je adresát (ADDR) řídícího slovesa *darovat* koreferující s aktorem (ACT) tohoto slovesa).

  VALLEX se omezuje na zachycení případů, kdy zájmeno *se/si* zaplňuje pozici aktantu s akuzativní nebo dativní formou.[13]

Poznámka: Atribut reflexivity se netýká případů, kdy je morfém *se/si* součástí slovesného lemmatu (tyto případy jsou popsány v odd. 2.1) nebo kdy je *se/si* příznakem reciprocity (těm je věnován následující odd. 5.3).

5.3 Reciprocita

Reciprocitou se rozumí možnost vyjádření vztahu vzájemnosti mezi dvěma (či více) valenčními doplněními, přičemž vztah mezi těmito doplněními je symetrický (doplnění přitom splňují jisté sémantické podmínky), viz Karlík et al. (2002).

Pokud je do vztahu reciprocity zapojen aktor (ACT), užívá se reflexivní (zvratné) sloveso, reciproční doplnění se potom vyjadřují jako koordinované členy podmětu (*Petr a Marie se hádali*) nebo podmět plurálový (*přátelé se navštěvují*); reciprocita může být zdůrazněna příslovci *spolu*, *navzájem* apod.

Pokud do vztahu reciprocity není zapojen aktor (ACT), reciproční vztah typicky vyplývá z koordinace či plurálové formy doplnění (např. *seznámil je*, *seznámil Jana a Marii*), konstrukce může být opět zdůrazněna příslovci *spolu*, *navzájem* apod.

Možnost recipročního užití je ve *VALLEXu* vyznačena v atributu reciprocity (značka rcp), jehož hodnotou jsou dvojice, příp. trojice funktorů identifikující doplnění, která mohou vstupovat do vztahu reciprocity

[13] *VALLEX* tedy nepokrývá případy, kdy se zájmeno *se/si* může vyskytovat v jiném pádu (např. *praštil sebou o postel*) či v předložkové skupině (např. *ode dneška děláme na sebe.PAT*, *nechali si to u sebe.LOC*).

(např. ACT-ADDR pro *hádat se – neustále se spolu hádali*, ACT-ADDR-PAT pro *mluvit – mluví spolu o sobě (navzájem)*).

V případě odvozených reflexiv (viz odd. 2.1), která je možno klasifikovat jako inherentně reciproční varianty slovesa, typicky s obligatorním doplněním s formou s+7 (viz Panevová, 2007; Panevová – Mikulová, 2007), je ve *VALLEXu* uváděna reciprocita u nereflexivního i reflexivního lexému.

VALLEX se omezuje na zachycení případů reciprocity, do které vstupují aktanty a obligatorní volná doplnění.

5.4 Syntakticko-sémantické třídy

Část lexikálních jednotek (2 903 z celkového počtu 6 460, tedy přibližně 45 % všech lexikálních jednotek) má určenu *syntakticko-sémantickou třídu* (značka class). Tyto třídy byly budovány striktně ‚zdola nahoru' – seskupováním lexikálních jednotek s podobnými syntaktickými vlastnostmi, přičemž se přihlíželo k jejich sémantice. Zdůrazněme zde, že syntakticko-sémantické třídy jsou tvořeny jednotlivými lexikálními jednotkami, nikoliv celými lexémy – víceznačný lexém se tedy může vyskytovat v několika třídách.

Bylo vytvořeno následujících 22 syntakticko-sémantických tříd:

- appoint verb (23 LU), např. *nominovat, určovat* (ve smyslu *určovala své zástupce*), *ustanovovat, …*;
- cause motion (43 LU), např. *hýbat* (*hýbat pravou rukou*), *mávat, vrhat, …*;
- combining (96 LU), např. *míchat* (*míchat žloutky s moukou v těsto*), *přidávat, spojovat, …*;
- communication (364 LU), např. *číst, hovořit, nařizovat, …*;
- contact (115 LU), např. *dotýkat se, narážet, tisknout, …*;
- emission (22 LU), např. *pouštět* (ve smyslu *tričko pouštělo barvu*), *vysílat* (ve významu *vysílat signály*), …;
- exchange (177 LU), např. *dávat, dostávat, měnit, platit, pronajímat, …*;
- expansion (19 LU), např. *pronikat, šířit, …*;
- extent (20 LU), např. *činit* (ve smyslu *činí to 30 Kč*), *dosahovat, vycházet* (ve smyslu *boty vycházejí na tisíc korun*), …;
- change (318 LU), např. *budovat, klesat* (ve smyslu *teplota prudce klesala*), *proměňovat, růst, …*;
- intervention (10 LU), např. *zasahovat, mluvit* (*do toho nemůžu mluvit*), …;
- location (399 LU), např. *doplňovat* (*doplňovat zboží do regálu*), *nacházet, shromažďovat, …*;
- mental action (304 LU), např. *cítit se* (ve smyslu *cítit se dobře*), *jásat, mrzet, …*;
- modal verb[14] (15 LU), např. *dovést* (ve smyslu *dovede plavat*), *chtít, moci, smět, …*;
- motion (309 LU), např. *běžet, dorážet, hýbat se* (*Nehýbej se!*), …;
- perception (104 LU), např. *hledět, pamatovat, všímat si, …*;
- phase of action (80 LU), např. *končit* (*přednáška končí v 5 hodin*), *vrcholit, vznikat, …*;
- phase verb (76 LU), např. *iniciovat, končit* (*končit školu*), *najet* (ve smyslu *najeli aspoň 500 mil*), …;
- providing (51 LU), např. *naplnit* (ve smyslu *naplnit vanu vodou*), *vybavovat, …*;
- psych verb (83 LU), např. *klamat, potěšit*, (ve smyslu *potěšila ho dárkem, dárek ho potěšil*), …;

[14] Ve zpracování modálních sloves, která jsou na pomezí gramatiky a lexika, se *VALLEX* odchyluje od teorie FGP. Ve FGP jsou modální slovesa *muset, mít, chtít, hodlat, moci, dát se, smět, dovést* a *umět* zachycena pomocí gramatémů u významových sloves (nemají tedy valenční rámec). Naproti tomu ve *VALLEXu* jsou kvůli úplnosti a lexikální proměnlivosti pro modální význam těchto sloves vyčleněny LU, nejsou však zachyceny všechny jejich syntaktické zvláštnosti. Protože některá z těchto sloves jsou víceznačná (např. *mít* je modální v užití *Jan má připravit večeři*, ale plnovýznamové v užití *Jan má spoustu peněz*), mohou být popsána v několika LU.

- social interaction (86 LU), např. *potkávat se* (*potkává se s přáteli v baru*), *spojovat* (*spojím se s ním co nejdříve*), *souhlasit*, ...;
- transport (189 LU), např. *donášet*, *přemisťovat/přemístťovat*, *shrnovat*, ...

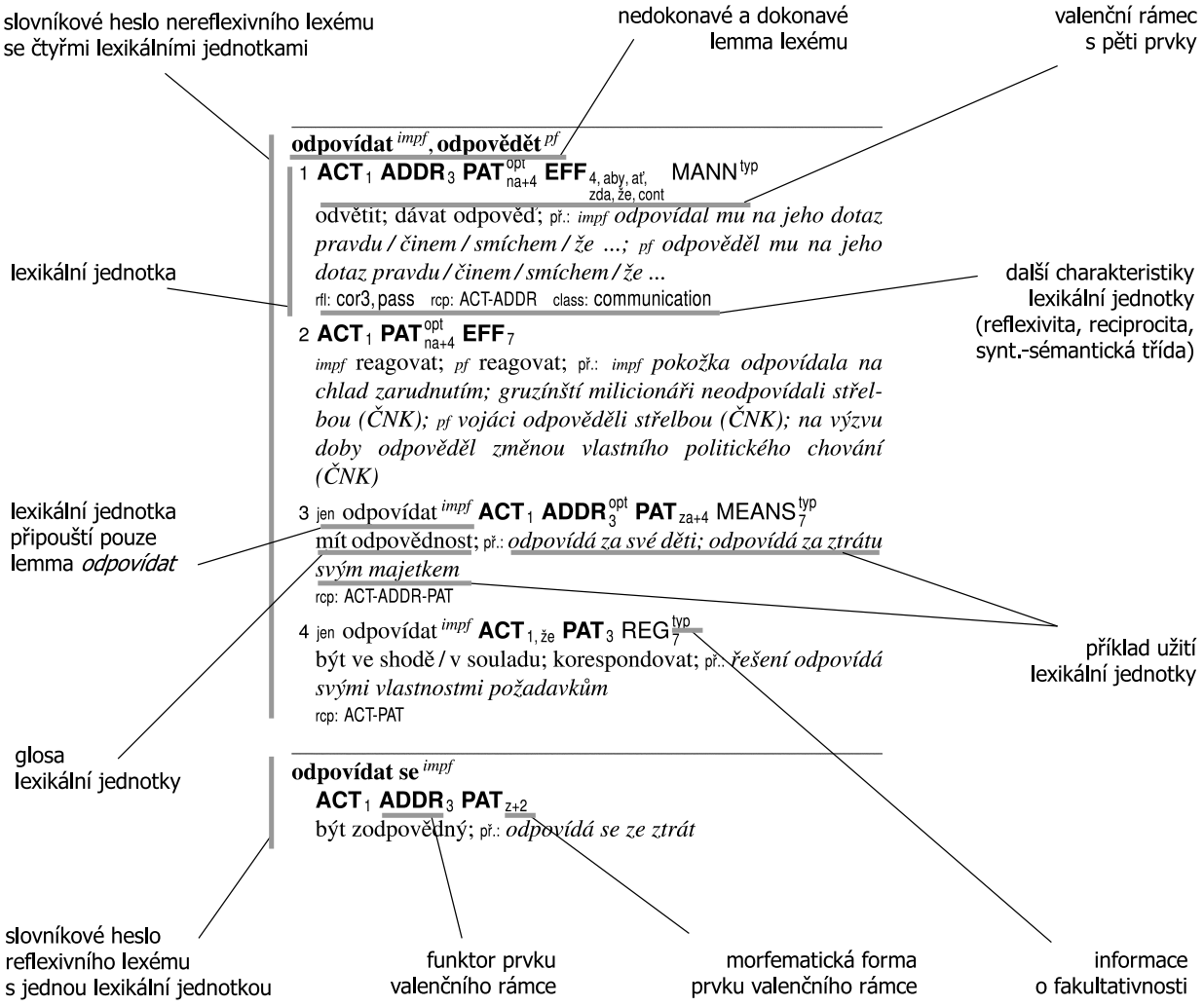
Upozorňujeme, že toto rozdělení lexikálních jednotek do syntakticko-sémantických tříd je pouze pracovní a nelze je považovat za klasifikaci splňující požadavky dobře definované ontologie. Je zřejmé, že takto hrubé rozdělení není syntakticky ani sémanticky homogenní, jde o základní vymezení skupin sloves, které je potřeba dále podrobně studovat. Motivací pro tuto předběžnou klasifikaci lexikálních jednotek byla skutečnost, že i takovéto pracovní třídění zachycuje vztahy mezi slovesy a díky tomu usnadňuje kontroly konzistence slovníku a dovoluje formulovat obecnější pozorování týkající se slovníkových dat.

5.5 Frazémy a idiomy

Při vytváření slovníku *VALLEX* byl kladen důraz především na úplné pokrytí primárních a obvyklých významů sloves. Zároveň bylo zpracováno mnoho lexikálních jednotek popisujících okrajová a idiomatická užití sloves; jejich pokrytí však není (a nemůže být) úplné. Takové lexikální jednotky jsou odlišeny značkou idiom za číslem lexikální jednotky.

Idiomatická užití sloves jsou taková ustálená užití, která jsou pracovně charakterizována buď podstatným posunem ve významu (vzhledem k primárnímu významu, např. *přišel o hodinky*), omezenou, obvykle velmi malou množinou možných lexikálních hodnot, kterých můžou jejich doplnění nabývat (např. *brát roha*, *mráz mi z toho běhal po zádech*), nebo jinými nepravidelnostmi a anomáliemi.

Poznámka: Metaforické užití slovesa – pokud nedošlo k jeho výrazné lexikalizaci – je obvykle pokryto lexikální jednotkou pro primární význam slovesa (například *po městě šla řeč, že se budeš stěhovat* je řazeno do lexikální jednotky slovesa *jít* popisující význam ‚pohybovat se po vlastních nohou; přemísťovat se chůzí‘).



Ukázka dvou slovníkových hesel

Podoba slovníkového hesla

Zde uvádíme pouze přehled struktury hesla pro usnadnění orientace (viz též obrázek na následující straně), jednotlivé pojmy jsou vysvětleny v následujícím oddíle.

Lemma – v záhlaví slovníkového hesla je uvedeno lemma (infinitivní tvar) reprezentující heslové sloveso, příp. seznam lemmat (v pořadí nedokonavé, dokonavé, iterativum), odd. 2.

Vid – jako horní index je u každého lemmatu uveden údaj o vidu (značky *impf*, *pf*, *iter*; případně následovány arabskou číslicí, pokud je slovníkové heslo reprezentováno více lemmaty se stejným videm), odd. 2.2.

Varianty – pokud má slovesné lemma varianty, jsou všechny varianty (oddělené lomítkem) uvedeny v záhlaví hesla, odd. 2.3.

Homografy – jsou rozlišeny římskou číslicí v dolním indexu u lemmatu, odd. 2.4.

Číslo LU – pokud má sloveso více významů, tzn. má více lexikálních jednotek (dále LU), je každá z nich označena arabskou číslicí, odd. 3.

Idiom – idiomatická užití jsou uvedena značkou idiom za číslem LU, odd. 5.5.

Omezení – pokud se daná LU vztahuje jen k některým lemmatům ze seznamu uvedeného v záhlaví, jsou za značkou jen uvedena všechna lemmata, ke kterým se tato LU vztahuje (s vyznačeným videm a případnými variantami a indexem pro homografy); omezení se neuvádí pro iterativa.

Valenční rámec – každá LU je popsána pomocí formálního zápisu rámce, který uvádí počet a typ (tzv. funktor) valenčních doplnění, jejich možná morfematická vyjádření (dolní index) a obligatornost (horní index – pokud není uveden, jde o obligatorní doplnění, značka opt zachycuje fakultativní doplnění, značka typ doplnění typické), odd. 4.

Glosa – každá LU je charakterizována glosou, která je na novém řádku za valenčním rámcem; pokud je heslové sloveso specifikováno několika lemmaty, jsou zde glosy pro všechna nedokonavá a dokonavá lemmata; glosy jsou uvedeny vždy údajem o vidu.

Příklad – každá LU obsahuje příklad užití uvedený značkou př.; pokud je heslové sloveso specifikováno několika lemmaty, jsou zde příklady pro všechna nedokonavá a dokonavá lemmata; příklady jsou vždy uvedeny údajem o vidu.

Doplňující syntaktické informace

- **kontrola** – slovesa kontroly mají za značkou control uveden funktor členu valenčního rámce tohoto slovesa, který koreferuje se subjektem infinitivu závislého slovesa (kontroluje ho), odd. 5.1;
- **reflexivita** – za značkou rfl jsou uvedeny možné syntaktické funkce morfému *se/si*, odd. 5.2;
- **reciprocita** – za značkou rcp jsou uvedeny dvojice, příp. trojice valenčních doplnění, která mohou vstupovat do vztahu reciprocity, odd. 5.3;
- **syntakticko-sémantická třída** – za značkou class je uvedena syntakticko-sémantická třída slovesa v daném významu, odd. 5.4.

Seznam literatury

APRESJAN, J. D. *Eksperimental´noje issledovanije semantiki russkogo glagola*. Moskva, Nauka, 1967.

BABKO-MALAYA, O. et al. Proposition Bank II: Delving Deeper. In MEYERS, A. (ed.) *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, s. 17–23, Boston, USA, 2004.

BLATNÁ, R. – ČERMÁK, F. (ed.). *Jak využívat Český národní korpus*. Praha, Nakladatelství Lidové noviny, 2005.

BOGUSLAVSKY, I. – IOMDIN, L. – SIZOV, V. Multilinguality in ETAP-3: Reuse of Lexical Resources. In *Proceedings of PostCOLING Workshop on Multilingual Linguistic Resources*, 2004.

BOND, F. – SHIRAI, S. Practical and Efficient Organization of a Large Valency Dictionary. In *Proceedings of the 4th Natural Language Processing Pacific*, Phuket, Thailand, 1997.

CHOMSKY, N. *Lectures on Government and Binding.* Dordrecht, Foris, 1981.

CINKOVÁ, S. From PropBank to EngValLex. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, s. 2170–2175. ELRA, 2006.

CRUSE, D. A. *Lexical Semantics*. Cambridge, Cambridge University Press, 1986.

DANEŠ, F. Větné členy obligatorní, potenciální a fakultativní. *Miscellanea Linguistica*. 1971, s. 131–138.

DANEŠ, F. *Věta a text.* Praha, Academia, 1985.

DANEŠ, F. The Sentence-Pattern Model of Syntax. In LUELSDORFF, P. A. (ed.) *The Prague School of Structural and Functional Linguistics*, s. 197–221. Philadelphia, John Benjamins Publishing Company, 1994.

DANEŠ, F. – HLAVSA, Z. *Větné vzorce v češtině*. Praha, Academia, 1987.

DANEŠ, F. – GREPL, M. – HLAVSA, Z. (ed.). *Mluvnice češtiny 3.* Praha, Academia, 1987.

DORR, B. J. et al. LCS Verb Database, Online Software Database of Lexical Conceptual Structures and Documentation. Technical report, University of Maryland, 2001.

DOWTY, D. *Word meaning and Montague grammar. The semantics of verbs and times in Generative Semantics and in Montague's PTQ: Synthese Language Library.* Dordrecht, Reidel, 1979.

ELLSWORTH, M. et al. PropBank, SALSA, and FrameNet: How Design Determines Product. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004), Workshop on Building Lexical Resources from Semantically Annotated Corpora*, Lisbon, 2004.

ERK, K. et al. Towards a Resource for Lexical Semantics: A Large German Corpus with Extensive Semantic Annotation. In *Proceedings of ACL-03*, Sapporo, Japan, 2003.

FILIPEC, J. – ČERMÁK, F. *Česká lexikologie*. Praha, Academia, 1985.

FILLMORE, C. J. The Case for Case. In BACH, E. – HARMS, R. T. (ed.) *Universals in Linguistic Theory*, s. 1–88. New York, Holt, Rinehart and Winston, 1968.

FILLMORE, C. J. Types of lexical information. In KIEFER, F. (ed.) *Studies in syntax and semantics*, s. 109–137. New York, Kluwer Academic Publishers, 1969.

FILLMORE, C. J. FrameNet and the Linking between Semantic and Syntactic Relations. In TSENG, S.-C. (ed.) *Proceedings of COLING 2002*, s. xxviii–xxxvi. Howard International House, 2002.

FILLMORE, C. J. – BAKER, C. – SATO, H. Seeing Argumens through Transparent Structures. In RODRÍGUEZ, M. G. – ARAUJO, C. P. S. (ed.) *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, s. 787–791. ELRA, 2002.

GREPL, M. – KARLÍK, P. *Skladba češtiny*. Olomouc, Votobia, 1998.

HAJIČ, J. Complex Corpus Annotation: The Prague Dependency Treebank. In ŠIMKOVÁ, M. (ed.) *Insight into Slovak and Czech Corpus Linguistics*, s. 54–73. Bratislava, Veda, 2005.

HAJIČ, J. et al. PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, 9, s. 57–68, 2003.

HAJIČ, J. et al. *Prague Dependency Treebank 2.0.* Philadelphia, PA, USA, Linguistic Data Consortium, 2006.

HAVRÁNEK, B. (ed.). *Slovník spisovného jazyka českého*. Praha, Academia, 1964.

HELBIG, G. – SCHENKEL, W. *Wörterbuch zur Valenz und Distribution deutscher Verben*. Leipzig, Bibliographisches Institut, 1969.

HLAVÁČKOVÁ, D. – HORÁK, A. VerbaLex – New Comprehensive Lexicon of Verb Valencies for Czech. In *Computer Treatment of Slavic and East European Languages*, s. 107–115, Bratislava, 2006.

JACKENDOFF, R. S. *Semantic Structures*. Cambridge, MIT Press, 1990.

KARLÍK, P. Hypotéza modifikované valenční teorie. *Slovo a slovesnost*. 2000, 61, s. 170–189.

KARLÍK, P. – NEKULA, M. – PLESKALOVÁ, J. (ed.). *Encyklopedický slovník češtiny*. Praha, Nakladatelství Lidové noviny, 2002.

KARLÍK, P. – NEKULA, M. – RUSÍNOVÁ, Z. (ed.). *Příruční mluvnice češtiny*. Praha, Nakladatelství Lidové noviny, 1996.

KIPPER, K. – DANG, H. T. – PALMER, M. Class-Based Construction of a Verb Lexicon. In *AAAI-2000 Seventeenth National Conference on Artificial Intelligence*, Austin, TX, 2000.

KIPPER-SCHULER, K. *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD thesis, Computer and Information Science Dept., University of Pennsylvania, Philadelphia, PA, 2005.

KIRCHMEIER-ANDERSEN, S. *Lexicon, Valency and the Pronominal Approach. An Application of the Pronominal Approach to Danish Verbs and Nouns.* PhD thesis, Odense Universitet, 1997.

KOMÁREK, M. et al. (ed.). *Mluvnice češtiny 2.* Praha, Academia, 1986.

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

LOPATKOVÁ, M. – PANEVOVÁ, J. Recent developments in the theory of valency in the light of the Prague Dependency Treebank. In ŠIMKOVÁ, M. (ed.) *Insight into Slovak and Czech Corpus Linguistics*, s. 83–92. Bratislava, Veda, 2006.

LOPATKOVÁ, M. – ŽABOKRTSKÝ, Z. – SKWARSKA, K. Valency Lexicon of Czech Verbs: Alternation-Based Model. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, 3, s. 1728–1733. ELRA, 2006.

MEJSTŘÍK, V. (ed.). *Slovník spisovné češtiny pro školu a veřejnost*. Praha, Academia, 2003.

MEL'ČUK, I. A. *Dependency Syntax: Theory and Practice*. Albany, State University of New York Press, 1988.

MEL'ČUK, I. A. Actants in semantics and syntax I: actants in semantics. *Linguistics*. 2004a, 42 (1), s. 1–66.

MEL'ČUK, I. A. Actants in semantics and syntax II: actants in syntax. *Linguistics*. 2004b, 42 (2), s. 247–291.

MEL'ČUK, I. A. – ZHOLKOVSKY, A. K. *Explanatory Combinatorial Dictionary of Modern Russian*. Vienna, Wiener Slawistischer Almanach, 1984.

MIKULOVÁ, M. et al. Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka. Technical Report TR-2005-28, ÚFAL MFF UK, Praha, 2005.

NIŽNÍKOVÁ, J. – SOKOLOVÁ, M. *Valenčný slovník slovenských slovies*. Prešov, Filozofická fakulta Prešovskej univerzity, 1998.

PALA, K. – ŠEVEČEK, P. Valence českých sloves. In *Sborník prací FFBU*, s. 41–54, Brno, 1997.

PANEVOVÁ, J. On Verbal Frames in Functional Generative Description. *The Prague Bulletin of Mathematical Linguistics*. 1974, 22, s. 3–40.

PANEVOVÁ, J. *Formy a funkce ve stavbě české věty*. Praha, Academia, 1980.

PANEVOVÁ, J. More Remarks on Control. *Prague Linguistic Circle Papers*. 1996, 2, s. 101–120.

PANEVOVÁ, J. Ještě k teorii valence. *Slovo a slovesnost*. 1998, 59, s. 1–13.

PANEVOVÁ, J. Znovu o reciprocitě. *Slovo a slovesnost*. 2007, 68, s. 91–100.

PANEVOVÁ, J. Some Issues of Syntax and Semantics of Verbal Modifications. In *Proceedings of the First International Conference on Meaning-Text Theory (MTT 2003)*, s. 139–146, Paris, 2003. Ecole Normale Supérieure.

PANEVOVÁ, J. Valency Frames and the Meaning of the Sentence. In LUELSDORFF, P. A. (ed.) *The Prague School of Structural and Functional Linguistics*, s. 223–243. Amsterdam, Philadelphia, John Benjamins Publishing Company, 1994.

PANEVOVÁ, J. – MIKULOVÁ, M. On Reciprocity. *The Prague Bulletin of Mathematical Linguistics*. 2007, 87, s. 27–40.

PANEVOVÁ, J. – ŘEZNÍČKOVÁ, V. K možnému pojetí všeobecnosti aktantu. In HLADKÁ, Z. – KARLÍK, P. (ed.) *Čeština - univerzália a specifika 3*, s. 139–146, 2001.

PANEVOVÁ, J. – SKOUMALOVÁ, H. Surface and Deep Cases. In *Proceedings of COLING 1992*, s. 885–889, Nantes, France, 1992.

PAULINY, E. *Štruktúra slovenského slovesa*. Bratislava, Slovenská akadémia vied a umení, 1943.

PETR, J. et al. (ed.). *Mluvnice češtiny 1*. Praha, Academia, 1986.

POLAŃSKI, K. (ed.). *Słownik syntaktyczno-generatywny czasowników polskich*. Wrocław, Wydawnictwo Polskiej Akademii Nauk, 1980–1992.

POPOVA, M. *Kratăk valenten rečnik na glagolite v săvremennia bălgarski knižoven ezik*. Sofia, Bulgarian Academy of Sciences Publishing House, 1987.

PUSTEJOVSKY, J. *The Generative Lexicon*. Cambridge, MIT Press, 1995.

RUPPENHOFER, J. et al. FrameNet II: Extended Theory and Practice (http://framenet.icsi.berkeley.edu/), 2006.

SGALL, P. Valence jako jádro jazykového systému. *Slovo a slovesnost*. 2006, 67, s. 163–178.

SGALL, P. *Generativní popis jazyka a česká deklinace*. Praha, Academia, 1967.

SGALL, P. Teorie valence a její formální zpracování. *Slovo a slovesnost*. 1998, 59, s. 15–29.

SGALL, P. – HAJIČOVÁ, E. – PANEVOVÁ, J. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht, Reidel, 1986.

SILNICKIJ, G. *Korreljacionnaja tipologija glagolnych sistem indoevropejskich i inostrukturnych jazykov*. Smolensk, Russia, Rossijskaja akademija nauk, Institut lingvističeskich issledovanij, 1999.

SKOUMALOVÁ, H. *Czech syntactic lexicon*. PhD thesis, Univerzita Karlova, Filozofická fakulta, 2001.

ŠMILAUER, V. *Novočeská skladba*. Praha, Nakladatel Ing. Mikuta, 1947.

ŠMILAUER, V. *Novočeská skladba*. Praha, SPN, 1966.

SVOZILOVÁ, N. – PROUZOVÁ, H. – JIRSOVÁ, A. *Slovesa pro praxi*. Praha, Academia, 1997.

SVOZILOVÁ, N. – PROUZOVÁ, H. – JIRSOVÁ, A. *Slovník slovesných, substantivních a adjektivních vazeb a spojení*. Praha, Academia, 2005.

TESNIÈRE, L. *Eléments de syntaxe structurale*. Paris, Librairie C. Klincksieck, 1959.

EYNDE, K. – MERTENS, P. La valence: l'approche pronominale et son application au lexique verbal. *French Language Studies*. 2003, s. 63–104.

ŽABOKRTSKÝ, Z. *Valency Lexicon of Czech Verbs*. PhD thesis, Charles University, Prague, 2005.

Chapter D

Formal Modeling of Natural Language: Valency as Core Syntactic Information

D.1 Modeling Syntax of Free Word-Order Languages: Dependency Analysis by Reduction

In Matoušek, V. – Mautner, P. – Pavelka, T. (eds.) *Proceedings of Text, Speech and Dialog International Conference, TSD 2005, 3658 / LNAI*, p. 140-147, Berlin Heidelberg, 2005. Springer-Verlag
(with co-authors M. Plátek and V. Kuboň)

Modeling Syntax of Free Word-Order Languages: Dependency Analysis by Reduction*

Markéta Lopatková[1], Martin Plátek[2], and Vladislav Kuboň[1]

[1] ÚFAL MFF UK, Praha
{lopatkova,vk}@ufal.mff.cuni.cz
[2] KTIML MFF UK, Praha
martin.platek@mff.cuni.cz

Abstract. This paper explains the principles of dependency analysis by reduction and its correspondence to the notions of dependency and dependency tree. The explanation is illustrated by examples from Czech, a language with a relatively high degree of word-order freedom. The paper sums up the basic features of methods of dependency syntax. The method serves as a basis for the verification (and explanation) of the adequacy of formal and computational models of those methods.

1 Introduction – Analysis by Reduction

It is common to describe the syntactic structure of sentences of English or other fixed word-order languages by phrase structure grammars. The description of the syntactic structure of Latin, Italian, German, Arabic, Czech, Russian or some other languages is more often based on approaches which are generally called dependency based. Both approaches are based on stepwise simplification of individual sentences, on the so-called analysis by reduction. However, the basic principles of the phrase-structure and dependency based analysis by reduction are substantially different. The phrase-structure based analysis (of fixed word-order languages) can be naturally modeled by the bottom-up analysis using phrase structure (Chomskian) grammars. This paper should help the reader to recognize that it is necessary to model the dependency analysis by reduction of languages with a high degree of word-order freedom differently. We try to explain explicitly the common basis of the methods for obtaining dependencies, presented in [3, 4, 7].

Unlike the artificial (programming) languages, the natural languages allow for an ambiguous interpretation. Instead of a complete formal grammar (of an artificial language), for natural languages we have at our disposal the ability of sentence analysis – we learn it at school, it is described by means of implicit rules in grammars of a given language.

The grammar textbooks are based on the presupposition that a human understands the meaning of a particular sentence before he starts to analyze it (let us cite from the 'Textbook of sentence analysis' (see [10]):"A correct analysis of a sentence is not

* This paper is a result of the project supported by the grant No. 1ET100300517. We would like to thank an anonymous reviewer for his valuable comments and recommendations.

V. Matoušek et al. (Eds.): TSD 2005, LNAI 3658, pp. 140–147, 2005.

possible without a precise understanding of that sentence, ... "). An automatic syntactic analysis (according to a formal grammar), on the other hand, neither does presuppose the sentence understanding, nor has it at its disposal. On the contrary, it is one of the first phases of the computational modeling of a sentence meaning.

What is actually the relationship between the sentence analysis and the analysis by reduction? In simple words, the sentence analysis is based on a more elementary ability to perform the analysis by reduction, i.e. to simplify gradually the analyzed sentences. The following simplified example illustrates the methodology of the dependency analysis by reduction.

Example 1. The sentence *'Studenti dělali těžkou zkoušku.' [Lit.: Students passed difficult exam.]* can be simplified (while preserving its syntactical correctness) in two ways (see also the scheme in Fig. 1) – by the deletion of the word form *studenti* or by the deletion of the word form *těžkou* (but not by the deletion of the word form *zkoušku* – the sentence *'*Studenti dělali těžkou.'* is not acceptable in a neutral context). In the second step we can remove the word form *těžkou* (in the first branch of the analysis) or the word form *studenti*, or even the word form *zkoušku* (in the second branch). In the last step we can delete the word form *zkoušku* (in the first branch), or the word form *studenti*.

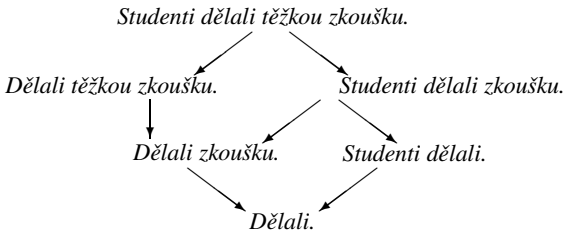


Fig. 1. The DAR scheme for the sentence *'The students passed a difficult exam.'*

The DAR scheme is closely related to a dependency tree, Fig. 2 shows the dependency tree for the sentence *Studenti dělali těžkou zkoušku.*

(i) A particular word depends on (modifies) another word from the sentence if it is possible to remove this modifying word (while the correctness of the sentence is preserved).

(ii) Two words can be removed stepwise in an arbitrary order if and only if they are mutually independent.

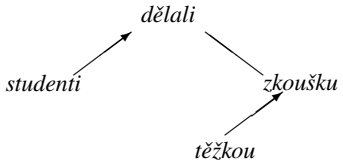


Fig. 2. The dependency tree for the sentence *'Studenti dělali těžkou zkoušku.'*

This example illustrates the way how we can obtain an information about dependencies (relationships between modified and modifying words in a sentence) using DAR. Let us stress the following fact: if taking correct Czech sentences with permuted word order, e.g. *'Těžkou zkoušku studenti dělali.'* or *'Těžkou dělali studenti zkoušku.'*, we get totally analogical reduction scheme as for the original sentence (the deleted words are identical in all steps of the reduction). This indicates that the dependency analysis by reduction allows to examine dependencies and word order independently. In other words, it provides a method for studying the degree of independence of the relationship between modified and modifying words in a sentence on its word order.

In this paper we concentrate on the description of rules for a dependency analysis by reduction of Czech, a language with a relatively high degree of word-order freedom, and on clarification of the relation between a dependency analysis by reduction and dependency sentence analysis.

The main reason for studying the analysis by reduction is the endeavor to gain a clear idea about its formal and computational modeling. Note that a formal model of analysis by reduction, restarting automata, is already intensively studied (see e.g. [3, 6]).

2 Dependency Analysis by Reduction

The **dependency analysis by reduction (DAR)** is based on stepwise simplification of a sentence – each step of DAR is represented by exactly one **reduction operation** which may be executed in two ways:

(i) by deleting at least one word of the input sentence, or

(ii) by replacing an (in general discontinuous) substring of a sentence by a shorter substring.

The possibility to apply certain reduction is restricted by the necessity to preserve some (at least the first one) of the following **DAR principles**:

(a) preservation of syntactical correctness of the sentence;

(b) preservation of lemmas and sets of morphological categories characterizing word forms that are not affected by the reduction operation;

(c) preservation of the meanings of words in the sentence (represented e.g. by valency frame[1], or by a suitable equivalent in some other language);

(d) preservation of the independence of the meaning of the sentence (the sentence has independent meaning if it does not necessarily invoke any further questions when uttered separately)[2].

With respect to a concrete task (e.g. for grammar checking) it is possible to relax these DAR principles; those which are not relaxed are then called **valid DAR principles** (e.g. in the example 1 we have relaxed the principle of preservation of the independence of sentence meaning).

If it is possible to apply a certain reduction in a certain step of DAR (preserving all valid principles), we talk about **admissible reduction**. By the application of all admis-

[1] The valency frame describes syntactic-semantic properties of a word, see e.g. [5].

[2] A sentence with independent meaning consists of a verb, all its semantically 'obligatory' modifications and (recursively) their 'obligatory' modifications, see [7].

sible reductions it is possible to get all **admissible simplifications** of a sentence being reduced.

We are going to use the term **DAR scheme (reduction scheme)** of a sentence of a given language for an oriented graph, whose nodes represent all admissible simplifications of a given sentence (including the original sentence) and whose edges correspond to all admissible reductions that can be always applied to a starting node of the edge and whose result is the admissible simplification of a sentence in its final node.

Example 2. The reduction scheme of the sentence *'Studenti dělali těžkou zkoušku.'* in Fig 1 illustrates the reductions of the type (i) – we delete at least one word of the input sentence in every step of the DAR whereas the possibility of branching captures the nondeterministic nature of the DAR. The reduction of the type (ii) is illustrated by possible simplification of the sentence *Kursem prošlo patnáct studentů. [Lit.: Course completed fifteen students.]*. Its reduction scheme is presented in Fig 3 (again, the principle (d) of the preservation of independence of meaning is relaxed).

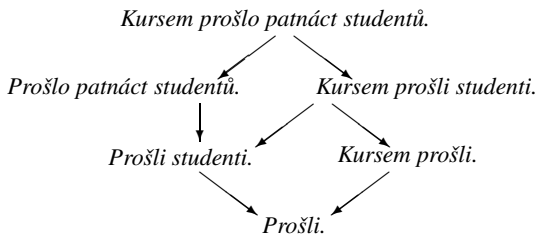


Fig. 3. The reduction scheme for the sentence *'Kursem prošlo patnáct studentů.'*

3 The Structure of Reduction and a Dependency Tree

The DAR scheme allows to introduce and classify various types of relationships. On the basis of these relationships we can define a structure of a sentence reduction.

Let us have a language L, a sentence $v \in L$, $v = v_1 v_2 ... v_m$, where $v_1, v_2, ..., v_m$ are the words, and a DAR scheme of the sentence v. We will say that the words v_i $i \in N, N \subseteq \{1, 2, ...m\}$ constitute a **reduction component**, if all words v_i are always removed at the same moment (i.e. in the DAR scheme all words v_i are removed in one step, which corresponds to a single edge in the scheme). We will say that the word v_i is **dependent (in the reduction)** on the word v_j, if the word v_i is deleted earlier than v_j in all branches of the DAR; the word v_j will be called a **governing (in the reduction)** word.

We will say that the words v_i and v_j are **independent on each other (with regard to the reduction)**, if they can be deleted in an arbitrary order (i.e. there is a DAR branch in which the word v_i is deleted earlier than the word v_j, and there is a DAR branch in which the word v_j is deleted earlier than the word v_i).

Based on the terms of dependency and component in the reduction we can define a reduction structure of a sentence, as it is illustrated in the following example.

Example 3. The reduction scheme of the sentence *'Studenti dělali těžkou zkoušku.' [Lit.: Students passed difficult exam.]* which preserves all DAR principles (including the principle (d) preservation of the independence of the meaning of the sentence) can be found on Fig. 4 – the verb *dělat* has two 'obligatory' modifications corresponding to a subject and a direct object, the noun *studenti* does not have obligatory modifications, therefore the sentence with independent meaning has a form *'Studenti dělali zkoušku.' [Lit.: Students passed exam.]*

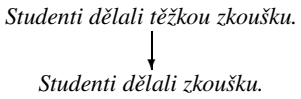


Fig. 4. The DAR scheme for the sentence *'Studenti dělali těžkou zkoušku.'* when applying the principle of preservation the independence of the sentence meaning.

The **reduction structure** can be captured by a diagram in which the nodes represent individual words from the sentence, the horizontal edges connect a reduction component (an edge always connects two neighboring words of a reduction component). The oblique edges reflect reduction dependencies; they are considered to be oriented from the dependent word (or from the whole reduction component) towards the governing word (or, again, towards the whole reduction component, if it is governing that particular word (component)). The linear order of nodes (left to right) captures the word-order (the order of words in the sentence). Fig. 5 shows the reduction structure representing the sentence *Studenti dělali těžkou zkoušku.*

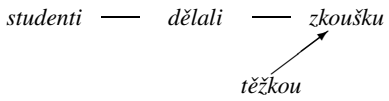


Fig. 5. The reduction structure for the sentence *'Studenti dělali těžkou zkoušku.'*

Traditionally, the structure of a (Czech) sentence is described by a dependency tree. Such a description is transparent and proper for sentences not complicated by coordinations, ellipses and by some marginal phenomena. The **dependency tree** is a structure that is a finite tree in the sense of a graph theory, and it has a root into which all paths are directed and whose nodes are totally (linearly left-to-right) ordered. The nodes represent the occurrences of word forms used in the sentence, the edges represent the relationship between a governing and a governed word (unit) in the sentence.

The only thing left to describe is how to get a dependency tree from a reduction structure. Reduction dependencies are easy, the respective edges characterize the relationship between the modifying and the modified word, the order of words in the sentence is preserved.

For reduction components it is necessary to find out which word from a given component will be considered as governing and which one will be dependent. For this pur-

pose it is necessary to introduce additional rules for individual linguistic phenomena, which are studied in more detail in the following section.

4 Reduction Relationships in a Natural Language

The formal typology of dependencies introduced in the previous section corresponds to a traditional linguistic classification – in this section we will try to describe this correspondence in more detail.

Let us suppose that the reader is familiar with basic linguistic notions such as subordination[3] (relation between modified sentence member and its modifying sentence member), complementation of verb/noun/adjective/adverb, inner participant (argument) and free modification (adjunct), obligatory and optional complementation. Description of these terms can be found e.g. in [9], [7] and [5].

Dependencies (in DAR) allow to model directly the optional free modifications – here it is possible to replace the whole pair by a modified word, a 'head' of the construction (without loosing the independence of meaning, the principle (d) of DAR). Thus we can capture the relationships like *těžká zkouška, jde pomalu, jde domů, přichází včas [Lit.: difficult exam, (she) walks slowly, (he) goes home, (he) comes in time]*. The governing word (in the reduction) corresponds to the modified word in the sentence, the dependent word (in the reduction) corresponds to the word which modifies it (see Fig. 6).

It remains to determine the governing and dependent member in those cases in which the modified or modifying member of this dependency consist of the whole reduction component, rather than of a single word.

(i) If the modifying member consists of the reduction component, then the dependent member is the governing word of this component (the remaining members of the component constitute a subtree with a root in this governing word).

(ii) If the modified sentence member consists of the reduction component, then the whole construction in general has ambiguous meaning (interesting examples for Czech can be found in [2]).

Fig. 6. Dependencies in DAR model free modifications.

Reduction components allow for modeling more complex relationships between word occurrences. These are either (a) morpho-syntactic relationships, or (b) syntactically-semantic relationships.

[3] The term of 'subordination' describes the language relationship, while the term of 'dependency' is reserved here for formal structures, by means of which language relationships are modeled.

(a) Reduction components describe so-called **formemes**, the units corresponding to individual sentence members – these are especially prepositional groups (as *na stole, vzhledem k okolnostem [Lit.: on table, with respect to circumstances]*) or complex verb forms (*přijel jsem, tiskne se [Lit.: (I) did arrive, (it) is being printed]*).

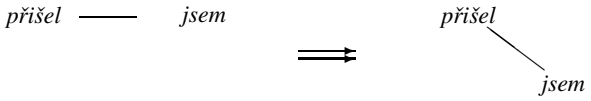


Fig. 7. A possible transformation of formemes into a dependency subtree.

In traditional linguistics each formeme constitutes one node of the diagram, or dependency tree describing syntactic structure of the sentence, see e.g. [10] or [9]. In these theories only the meaningful words (especially meaningful verbs, nouns, adjectives and adverbs) are represented by independent nodes. However, for many practically oriented tasks (e.g. grammar-checking, building of a syntactically annotated corpus) it is appropriate to represent each word of a sentence by its own node. In order to preserve the traditional data type of the dependency tree it is necessary to specify additional rules on the basis of which even the reduction components can be transformed into subtrees, i.e. it is necessary to specify which word from the formeme will be considered governing and which one will be dependent. Such rules are usually of a technical nature and they can differ in individual projects (Fig. 7 shows the solution adopted in [1]).

(b) The second type of relationships modeled by reduction components are syntactically-semantic relationships. These are especially **valency relationships** – the relationships of a verb, noun, adjective or adverb and its obligatory valency complementation(s) (as e.g.*studenti dělali zkoušku, Petr dal Pavlovi dárek, začátek přednášky [Lit.: students passed exam, Petr gave Pavel gift, beginning (of) lecture]*). These constructions cannot be replaced by a single word, the 'head' of the construction, without loosing the independence of meaning, DAR principle (d).

Traditional linguistics captures the valency relationships using dependency tree (see [9] and [10]). The theoretical criterion for the determination of modified and modifying sentence member, the principle of analogy in the layer of word classes is discussed in [9] – the verb is considered as a modified word (as an analogy to verbs without obligatory complementations), the verb complementations are the modifying words; similarly for nouns, adjectives, adverbs and their complementations. This principle of analogy is also adopted for determining the governing word during the transformation of reduction structure to a dependency tree: the verb is considered as a governing word, the verb complementations are its dependent words; similarly for nouns, adjectives, adverbs.

Let us note that the analogy principle can be simply substituted by a relaxation of the condition (d) preserving the independence of meaning of DAR.

Concluding Remarks

The DAR allows to formulate the relationship of basic syntactic phenomena: a dependency and a word order. This approach is indispensable especially for modeling

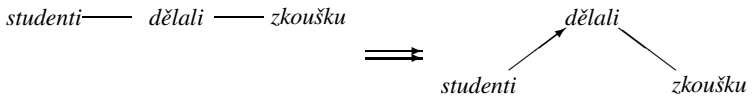


Fig. 8. The transformation of valency relationships into a dependency subtree.

the syntactic structure of languages with a free word-order, where the dependency and word-order are very loosely related and where they are also related in a different manner from language to language (let us compare this situation with English, where the dependencies are determined (mainly) by a very strict word-order).

The paper shows that the dependencies can be derived from two different, not overlapping, simply observable and language independent phenomena: from the reduction dependency and from reduction components. It also points out that the (Czech) traditional linguistic taxonomy of language phenomena corresponds to this division. We have mentioned the formal model of analysis by reduction, restarting automata. We have thus outlined one important step how to pass the observations about dependencies from traditional linguistics into the formal terms suitable for computer linguistics.

References

1. Hajič, J.: Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In: Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová (ed. Hajičová, E.), Karolinum, Prague, pp. 106-132 (1998)
2. Holan, T., Kuboò, V., Oliva, K., Plátek, M.: On Complexity of Word Order. In: Les grammaires de dépendance - Traitement automatique des langues (TAL), Vol. 41, No. 1 (q.ed. Kahane, S.), pp. 273-300 (2000)
3. Jančar, P, Mráz, F., Plátek, M., Vogel, J.: On Monotonic Automata with a Restart Operation. Journal of Automata, Languages and Combinatorics, Vol. 4, No. 4, pp. 287-311 (1999)
4. Kunze, J.: Abhängigkeitsgrammatik. Volume XII of Studia Grammatica, Akademie Verlag, Berlin (1975)
5. Lopatková, M.: Valency in the Prague Dependency Treebank: Building the Valency Lexicon. In: PBML 79-80, pp. 37-59 (2003)
6. Otto, F.: Restarting Automata and their Relations to the Chomsky Hierarchy. In: Developments in Language Theory, Proceedings of DLT'2003 (eds. Esik, Z., Fülöp, Z.), LNCS 2710, Springer, Berlin (2003)
7. Panevová, J.: Formy a funkce ve stavbě české věty. Academia, Praha (1980)
8. Plátek, M., Lopatková, M., Oliva, K.: Restarting Automata: Motivations and Applications. In: Proceedings of the workshop "Petrinetze" (ed. Holzer, M.), Technische Universität München, pp. 90-96 (2003)
9. Sgall, P., Hajičová, E., Panevová, J.: The Meaning of the Sentence in Its Semantic and Pragmatic Aspects (ed. Mey, J.), Dordrecht:Reidel and Prague:Academia (1986)
10. Šmilauer, V.: Učebnice větného rozboru. Skripta FF UK, SPN, Praha (1958)

D.2 Functional Generative Description, Restarting Automata and Analysis by Reduction

In Marušič, F. – Žaucer, R. (eds.) *Studies in Formal Slavic Linguistics. Contributions from Formal Description of Slavic Languages 6.5., Vol. 19 / Linguistik International*, p. 173-190. Frankfurt am Main: Peter Lang Publishing Group, 2008
(with co-authors M. Plátek and P. Sgall)

FDSL 6.5, 173-190

Markéta Lopatková, Martin Plátek, and Petr Sgall

Functional Generative Description, Restarting Automata and Analysis by Reduction*

1. Introduction

Functional Generative Description (FGD is a dependency based system for Czech, whose beginnings date back to the 1960s (see esp. Sgall *et al.* 1969, Sgall *et al.* 1986). FGD may be of some interest for the description of most Slavic languages, since it is adapted to treat a high degree of *free word order*. It not only specifies surface structures of the given sentences, but also translates them into their underlying representations. These representations (called tectogrammatical representations, denoted TRs) are intended as an appropriate input for a procedure of semantico-pragmatic interpretation in the sense of intensional semantics (see Hajičová *et al.* 1998). Since TRs are, at least in principle, disambiguated, it is possible to understand them as rendering linguistic (literal) meaning (whereas figurative meaning, specification of reference and other aspects belong to individual steps of the interpretation).

FGD has been implemented as a generative procedure by a sequential composition of pushdown automata (see Sgall *et al.* 1969, Plátek *et al.* 1978). Lately, as documented e.g. in Petkevič (1995), we have been interested in the formalization of FGD designed in a declarative way. In the present paper we want to formulate a formal framework for the procedure of checking the appropriateness and completeness of a description of a language in the context of FGD. The first step in this direction was introduced in Plátek (1982), where the formalization by a sequence of translation schemes is interpreted as an analytical system, and as a generative system as well. Moreover, requirements for a formal system describing a natural language *L* have been formulated – such a system should capture the following issues:

- The set of correct sentences of the language *L*, denoted by *LC*.
- The formal language *LM* representing all possible tectogrammatical representations (TRs) of sentences in *L*.
- The relation *SH* between *LC* and *LM* describing the ambiguity and the synonymy of *L*.

* This paper is a result of the project supported by the grants No. 1ET100300517 and MSM0021620838. The extended version is prepared for The Prague Bulletin of Mathematical Linguistics.

- The set of the correct structural descriptions *SD* representing in a structural way all possible TRs of sentences in *L* as dependency-based structures (*dependency trees*).

The object of the present paper concerns the foundations of a *reduction system* which is more complex than a reduction system for a (shallow) syntactic analyzer, since it provides not only the possibility of checking the well-formedness of the (surface) analysis of a sentence, but its underlying (tectogrammatical in terms of FGD) representation as well. Such a reduction system makes it possible to define formally the *analysis* as well as the *synthesis* of a sentence.

We propose here a new formal frame for checking FGD linguistic descriptions, based on *restarting automata*, see e.g. Otto (2006), Messerschmidt *et al.* (2006). We fully consider the first three requirements, i.e., *LC*, *LM* and *SH*. The fourth one is not formally treated here.

The main contribution of the new approach consists in the fact that it mirrors straightforwardly the so-called *(multi-level) analysis by reduction*, an implicit method used for linguistic research. Analysis by reduction consists of stepwise correct reductions of the sentence; roughly speaking, the input sentence is simplified until the so-called *core predicative structure* of the sentence is reached. It allows for obtaining (in)dependencies by the *correct reductions* of Czech sentences as well as for describing properly the complex word-order variants of a language with a high degree of `free' word order (see Lopatková *et al.* 2005). During the analysis by reduction, a (disambiguated) input string is processed, i.e., a string of tokens (word forms and punctuation marks) enriched with metalanguage categories from all linguistic layers encoded in the sentence.

In Section 2., we provides a brief characterization of analysis by reduction (subsection 2.1.) and then we address two basic linguistic phenomena, dependency (subsection 2.2.) and word order (2.3.), and show the process of the analysis by reduction on examples from Czech.

Now, let us briefly describe the type of restarting automaton that we use for modeling analysis by reduction for FGD (see Section 3). A *4-LRL-automaton* M_{FGD} is a non-deterministic machine with a finite-state control Q, a finite characteristic vocabulary Σ (see below), and a head (window of size 1) that works on a flexible tape. The automaton M_{FGD} performs:

- *move-right* and *move-left steps*, which change the state of M_{FGD} and shift the window one position to the right or to the left, respectively,
- *delete steps*, which delete the content of the window, thus shortening the tape, change the state, and shift the window to the right neighbor of the symbol deleted.

At the right end of the tape, M_{FGD} either halts and *accepts* the input sentence, or it halts and *rejects*, or it *restarts*, that is, it places its window over

the left end of the tape and reenters the initial state. It is required that before the first restart step and also between any two restart steps, M_{FGD} executes at least one delete operation.

The *4-LRL*-automata can be also represented by a final set of so called metarules (see Messerschmidt *et al.* 2006), a declarative way of representation, which seems to be a very promising tool for natural language description.

The basic notion related to M_{FGD} is the notion of the language accepted by M_{FGD}, so called *characteristic language* $L_{\Sigma}(M_{FGD})$. In our approach, it is considered as a language that consists of all sentences from the surface language *LC* over alphabet Σ_0 enriched with metalanguage information from Σ_1, Σ_2, Σ_3. The tectogrammatical language *LM* as well as the relation *SH* can be extracted from $L_{\Sigma}(M_{FGD})$.

In order to model the analysis by reduction for FGD, the *4-LRL*-automaton M_{FGD} works with a complex characteristic vocabulary Σ that is composed from (sub)vocabularies Σ_0, … Σ_3. Each subvocabulary Σ_i represents the corresponding layer of language description in FGD, namely:

- Σ_0 is the set of Czech written *word-forms* and *punctuation marks* (tokens in the sequel), it is the vocabulary for the language *LC* from the request 1 above;
- Σ_1 represents the *morphemic layer* of FGD, namely morphological lemma and tag
- for each token;
- Σ_2 describes surface syntactic functions (as e.g., Subject, Object, Predicate);[1]
- Σ_3 is the vocabulary of the *tectogrammatical layer* of FGD describing esp. `deep' roles, valency frame for frame evoking words, and meaning of morphological categories.

That means that the automaton has an access to all the information encoded in the processed sentence (as well as a human reader/linguist has all the information for his/her analysis).

M_{FGD} was introduced with no ambitions to model directly the procedure of the sentence-generating in the human mind or of the procedure of understanding performed in the human mind. On the other hand, it has a straightforward ambition to model the observable behavior of a linguist performing *analysis by reduction* of Czech sentences on the blackboard or on a sheet of paper.

1 Note that the layer of surface syntax does not correspond to any layer present in the theoretical specification of FGD, but rather to the auxiliary 'analytical' layer of the Prague Dependency Treebank, see Mikulová *et al.* (2005), which is technically useful for a maximal articulation of the process of analysis.

2. Analysis by Reduction for FGD

In this section we focus on the analysis by reduction for Functional Generative Description. After a brief characterization of analysis by reduction (subsection 2.1.), we address two basic linguistic phenomena, dependency (subsection 2.2.) and word order (2.3.), and illustrate the process of the analysis by reduction on examples from Czech.

2.1. Analysis by Reduction

The analysis by reduction makes it possible to formulate the relationship between dependency and word order (see also Lopatková *et al.* 2005). This approach is indispensable especially for modeling the syntactic structure of languages with a high degree of 'free' word order, where the dependency (predicate-argument) structure and word order are very loosely related. The restarting automaton M_{FGD} that models analysis by reduction for FGD is specified in detail in the Section 3.

The *analysis by reduction* is based on a stepwise simplification of a sentence – each step of analysis by reduction consists of deleting at least one word of the input sentence (see Lopatková *et al.* 2005 for more details).[2] The following principles must be satisfied:

- preservation of syntactic correctness of the sentence;
- preservation of the lemmas and sets of morphological categories;
- preservation of the meanings/senses of the words in the sentence (represented e.g. as an entry in a (valency) lexicon);
- preservation of the 'completeness' of the sentence (in this text only valency complementations (i.e., its arguments/inner participants and those of its adjuncts/free modifications that are obligatory) of frame evoking lexical items must be preserved).

The analysis by reduction works on a sentence (string of tokens) enriched with metalanguage categories from all the layers of FGD – in addition to word forms and punctuation marks, it embraces also morphological, surface and tectogrammatical information.

The input sentence is simplified until the so called *core predicative structure* of the sentence is reached. The core predicative structure consists of:

- the governing verb (predicate) of an independent verbal clause and its valency complementations, or
- the governing noun of an independent nominative clause and its valency complementations, e.g., *Názory čtenářů.* [Readers' opinions.], or

2 Here we work only with the deleting operation whereas in Lopatková *et al.* (2005) the rewriting operation is also presupposed.

– the governing word of an independent vocative clause, e.g., *Jano!* [Jane!], or
– the governing node of an independent interjectional clause, e.g., *Pozor!* [Attention!].

2.2. Processing dependencies

Czech is a language with a high degree of so-called free word order. Naturally, (surface) sentences with permuted word order are not totally synonymous (as the word order primarily reflects the topic-focus articulation in Czech), but their grammaticality may not be affected and the dependency relations (as binary relations between governing and dependent lexical items) may be preserved regardless of the word order changes. This means that the identification of a governing lexical item and its particular complementations is not based primarily on their position in the sentence but rather on the possible order of their reductions.

There are two ways of processing dependencies during the analysis by reduction.

– Free modifications (i.e., adjuncts) that do not satisfy valency requirements of any lexical item in the sentence are deleted one after another, in an arbitrary order (sentence (1)).
– The so called reduction components (formed by words that must be reduced together to avoid non-grammaticality, i.e., incompleteness of tectogrammatical representation)[3] are processed 'en bloc' depending on their function in the sentence:
  - Either all members of the reduction component are reduced – this step is applied if the 'head' of the reduction component does not fulfill any valency requirements of any lexical item in the sentence (see sentence (2) below where the whole component represents optional free modification).
  - Or (if the 'head' of the reduction component satisfies the valency frame of some lexical item):
    (i) the item representing the 'head' is simplified – all the symbols apart from the functor[4] are deleted; the result of such a simplification can be understood as a zero lexical realization of the respective item, see sentence (3) below;

3 Typically, a reduction component is composed of a frame evoking lexical item together with its valency complementations, see Lopatková *et al.* (2005). Let us stress here that a reduction component may constitute a discontinuous string.

4 A functor is the label for syntactico-semantic relation holding between the respective item and its governing lexical item.

(ii) the complementation(s) of the `head' of the reduction component is/are deleted.

Convention: For the sake of clarity we have adopted the following conventions for displaying examples:

– Each column contains a symbol from one part of the (partitioned) vocabulary, that means information on one layer of FGD:[5]
  - the first column contains tokens,
  - the second column contains morphological lemmas (m-lemmas) and morphemic values (i.e., morphological categories),
  - the third column contains (surface) syntactic functions,
  - for autosemantic words,[6] the fourth column contains tectogrammatical lemmas (t-lemmas), functors, frame identifiers and other tectogrammatical categories (so called grammatemes).
– Each individual token and its metalanguage categories are located:
  - in one line if its surface word order position agrees with the deep word order (i.e., word order at the tectogrammatical layer), or the token has no tectogrammatical representation (i.e., it is not an autosemantic word);
  - in two lines if its surface word order position disagrees with the deep word order:
    (i) one line embraces the token, its m-lemma and morphemic values as well as its (surface) syntactic function, and
    (ii) the other line contains relevant tectogrammatical information (for autosemantic words).
– The top-down ordering of lines reflects the word order on the respective layer.

Such a two-dimensional convention allows for revealing both (i) a representation of a whole sentence on particular layers (individual columns for particular layers), including relevant word order (columns 1, 2, 3 reflect the surface word order whereas column 4 is organized according to deep word order), and (ii) information relevant for individual tokens (rows).

Let us illustrate the processing of dependencies on the examples.

Example:

(1) *Včera přišel domů pozdě.*
yesterday came home late
"Yesterday he came home late."

5 The standard notation used in the Prague Dependency Treebank is used, see Hajič (2005).

6 Function words have just functors or grammatemes as their tectogrammatical correlates that are assigned to their governing autosemantic words.

The analysis by reduction starts with the input string specified in Fig. 1. (see the convention above; the metalanguage categories are explained e.g. in Hajič 2005).[7]

Včera	*m-včera*.Dg- - -	Adv	*t-včera*.TWHEN [on].ACT
přišel	*m-přijít*.VpYS-	Pred	*t- přijít*.PRED.Framel.ind-ant
domů	*m-domů*.Db- - -	Adv	*t-domů*.DIR3
pozdě	*m-pozdě*.Dg- - -	Adv	*t-pozdě*.TWHEN
.	..Z: - - -	AuxK	

Fig. 1. The input string for sentence (1).

It is obvious that an item of TR (an autosemantic word, see for Note 6) can have zero surface lexical realization (e.g., actor, ACT need not be realized, as Czech is a pro-drop language – the corresponding item is restored in the TR; also different kinds of ellipsis are possible). On the other hand, several word forms can constitute a single item of TR (as e.g., a prepositional group in sentence (2)).

Let us point out the difference between the two types of free modifications in the sentence, namely DIR3 (direction `to_where') and TWHEN (temporal relation `when'): (i) whereas the valency complementation of direction DIR3 is considered to be obligatory for the verb *přijít* [to come] (the speaker as well as the listener must know this, see the dialogue test proposed in Panevová 1974) and thus fills the relevant slot of the valency frame of the verb (here marked by the label Frame1), (ii) the temporal relation TWHEN is an optional free modification (not belonging to the valency frame Frame1).

(2 steps) →

			[on].ACT
přišel	*m-přijít*.VpYS-	Pred	*t- přijít*.PRED.Framel.ind-ant
domů	*m-domů*.Db- - -	Adv	*t-domů*.DIR3
.	..Z: - - -	AuxK	

Fig. 2. The reduced string – a core predicative structure for sentence (2).

The first step of analysis by reduction consists in the deletion of one of the optional free modifications *včera* [yesterday] or *pozdě* [late].[8] These free

7 We leave aside the problems of word order – this domain is briefly addressed in the following subsection.

8 More precisely, the tokens as well as all the metalanguage categories relevant for the particular lexical item are reduced, similarly in the sequel.

modifications may be reduced in an arbitrary order, they are mutually independent (see Lopatková *et al.* 2005). These reduction steps result in the string in Fig. 2.

Now, the sentence contains only one reduction component constituted by the finite verb and its valency complementations, i.e., its actor (expressed by a zero form of the pronoun) and its obligatory free modification DIR3 'to_where', *[on] přišel domů* [(he) came home]. This is a core predicative structure, thus the reduction ends successfully.

Example:

(2) *Petr včera přišel do školy, kterou loni postavil*
Peter yesterday came to school which last_year built
minulý starosta.
previous major
"Yesterday Peter came to the school which was built last year by the previous mayor."

This example shows the reduction of the whole reduction component that consists of a dependent clause. The input string looks as in Fig. 3.

Petr	*m-Petr*.NNMS1	Sb	*t-Petr*.ACT
včera	*m-včera*.Dg- - -	Adv	*t-včera*.TWHEN
přišel	*m-přijít*.VpYS-	Pred	*t- přijít*.PRED.Framel.ind-ant
do	*m-do*.RR- - 2	AuxP	
školy	*m-škola*.NNFS2	Adv	*t-škola*.DIR3.basic
,	„.Z: - - -	AuxK	
kterou	*m-který*.P4FS4	Obj	*t-který*.PAT
loni	*m-loni*.Db- - -	Adv	*t-loni*.TWHEN
postavil	*m-postavit*.VpYS-	Atr	*t-postavit*.RSTR.Frame2.ind-ant
minulý	*m-minulý*.AAMS1	Atr	
starosta	*m-starosta*.NNMS1	Sb	*t-starosta*.ACT *t-minulý*.RSTR
.	..Z: - - -	AuxK	

Fig. 3. The input string for sentence (2).

In the first three steps, the three optional free modifications *včera*, *loni* and *minulý* [yesterday, last_year, previous] are deleted in arbitrary order.

Next, the whole component *kterou postavil starosta* [which the mayor built] consisting of the verb and its valency complementations is to be processed. As this component represents an optional adnominal free modification RSTR, it can be simply deleted without the loss of completeness.

After this step, only one reduction component *Petr přišel do školy* [Peter came to school] remains, which constitute a core predicative structure – the analysis by reduction ends successfully.

Example:

(3) *Petr pomáhal Marii uklízet zahradu.*
Peter helped Mary to clean garden
"Peter helped Mary to clean the garden."

In this example there is a valency complementation realized as an infinitive form of the verb *uklízet* [to clean] and its two valency complementations, *[ona]* [she] (non-expressed) and *zahradu* [garden].[9]

In order to obtain the core predicative structure, the following simplification of the reduction component is used: (i) the complementations *[ona]* [she] and *zahradu* [garden] of the head verb *uklízet* [to clean] are deleted and (ii) the word form *uklízet* [to clean] and all the categories relevant to this word form apart from its functor (here PAT, patient) are deleted – such a simplified item represents a (saturated) lexical item with zero morphemic form (and thus, the valency requirements remain satisfied).

This step results in the core predicative structure.

2.3. Word Order

A large effort has been devoted to clearing up the role of word order in so called free-word order languages, see e.g. Hajičová *et al.* (1998), Holan *et al.* (2000), Havelka (2005), and Hajičová (2006) for some of the most recent contributions for Czech.

Let us recall two basic principles for the tectogrammatical representation of FGD (see esp. Sgall *et al.* 1986 and Hajičová *et al.* 1998):

- The word order in TR (deep word order) reflects the topic-focus articulation – it corresponds to the scale of communicative dynamism (thus it may differ from the surface word order).
- The theoretical research assumes the validity of the principle of projectivity for TRs.

These two principles have important consequences for the analysis by reduction that models the transition from surface form of a sentence to its TR – the surface word order must be modified in order to obtain the deep word order (example (4)). This holds particularly for sentences with non-projective surface

9 We leave aside the relation of control, i.e., a specific type of grammatical coreference between a complementation of a governing node and (non-expressed) subject of the infinitive verb.

structure (example (5)). It implies that the sentence representation must in general reflect two word orders, the surface and the deep one. Let us repeat here the adopted convention of displaying examples, particularly that for word order – whereas columns 1, 2, 3 depict surface word order, column 4, reflecting tecto-grammatical representation, reveals the deep word order.

Example: (see Mikulová *et al.* (2006), Sectiom 10.3.1.)

(4) *Černý kocour se napil ze své misky.*
black tomcat *refl* drunk from his bowl
"The black tomcat drank from its bowl."

Let us concentrate here on the topic focus articulation (see esp. Hajičová *et al.* 1998 and the writings quoted there).

According to Mikulová *et al.* (2006), the most general guideline of representing deep word order in TR is the placing of nodes representing contextually bound expressions to the left from their governing node and the placing of nodes representing contextually non-bound expressions to the right from their governing node. The contextual boundness is described in the attribute `tfa', the values `c' (contrastive topic), `t' (contextually bound) and `f' (contextually non-bound) belong to the metalanguage categories in the tecto-grammatical vocabulary. The input string for analysis is in Fig. 4 (the last category in the fourth column, divided by `_', reflects tfa).

Černý	*m-černý*.NNMS1	Atr	
kocour	*m-kocour*.NNMS1	Sb	*t-kocour*.ACT_t
			t-černý.RSTR_f
			[Gen].PAT_t
se	*m-se*.P7-X4	AuxR	
napil	*m-napít*.VpYS-	Pred	*t-napít_se*.PRED.Frame5_f
ze	*m-z*.RV- - 2	AuxP	
své	*m-svůj*.P8FS2	Atr	[PersPron].APP_t
misky	*m-miska*.NNFS2	Adv	*t-miska*.DIR1.basic_f
.	..Z: - - -	AuxK	

Fig. 4. The input string for sentence (4).

The actor, ACT *kocour*_t [tomcat] is contextually bound and it appears to the left of its governing verb *napil_se*_f [drank] in the surface; the contextually non-bound DIR1 complementation *misky*_f [bowl] is to the right of its governing verb; and the contextually bound *svůj*_t [his] is to the left from its governing word *miska*_f [bowl] as well – the surface word order agrees in these cases with the deep word order.

On the other hand, the modification *černý*_f [black] is contextually non-bound and it stands before its (bound) governing word *kocour*_t [tomcat] – here the surface word order disagrees with the deep word order. This is the reason why the ordering in the last column (with the tectogrammatical representation) does not replicate the ordering of other columns – the contextually bound modification *černý*_f [black] appears at the second position in the TR of the sentence (just behind the governing item *kocour*_t [tomcat]).

Now, the reduction phase can start, i.e., a stepwise simplification of the sentence according to the principles of analysis by reduction, during which the dependencies are treated and the core predicative structure is obtained, as it is described in the previous subsection.

Example: (see Sgall *et al.* 1986, p. 241)

(5) *Karla plánujeme poslat na rok do Anglie.*
Charles (we) plan to_send for year to England
"Charles we are planning to send for a year to England."
≈ As for Charles, we are planning to send him for a year to England.

The proper noun *Karla*_c [Charles], which is the contrastive topic of a sentence (tfa = `c'), is moved away from its governing verb *poslat*_f [to send], which causes a non-projectivity in the surface structure. The theoretical assumption of projectivity of TRs requires a different deep order – the corresponding item *t-Charles*.PAT_c in TR is situated just before its governing item *t-poslat*.PRED.Frame1_f [to send].

The analysis by reduction has the input string as in Fig. 5.

Karla	*m-Karel*.NNMS4	Obj	[my].ACT_t
plánujeme	*m-plánovat*.VB-P-	Pred	*t-plánovat*.PRED.Frame6.ind-sim_f *t-Karel*.PAT_c [my].ACT_t
poslat	*m-poslat*.Vf- - -	Obj	*t-poslat*.PAT.Frame7_f
na	*m-na*.RR- - 4	AuxP	
rok	*m-rok*.NNIS4	Adv	*t-rok*.THL_f
do	*m-do*.RR- - 2	AuxP	
Anglie	*m-Anglie*.NNFS2	Adv	*t-Anglie*.DIR3.basic_f
.	..Z: - - -	AuxK	

Fig. 5. The input string for sentence (5).

Now, the reduction phase treating the dependencies can start.

3. The *4-LRL*-automata

In this section, the formal model for analysis by reduction for FGD is proposed. We use here the standard way of presentation from the theory of automata (our remarks should hopefully help readers not quite familiar with that kind of presentation). This section is partitioned into two subsections. The first one introduces *sRL-automata* – the basic models of restarting automata we will be dealing with. The important notion of metarules is introduced here; they serve for a more transparent, more declarative description of restarting automata.

The second subsection introduces *4-LRL-automata* as a special case of *sRL*-automata. A four-level *analysis by reduction system*, which is an algebraic representation of analysis by reduction, and the formal languages which represent the individual layers of FGD are introduced here, namely the languages of the first and the last level that correspond to the surface language *LC* and to the tectogrammatical language *LM* from Section 1. Further, the *characteristic relation SH(M)* is introduced.

Finally, the *SH-synthesis*, which models FGD as a generative device and specifies the generative ability of FGD, and *SH-analysis*, which fulfills the task of syntactico-semantic analysis of FGD, are introduced here step by step.

3.1. The *t-sRL*-Automaton

Here we describe in short the type of restarting automaton we will be dealing with. The subsection is an adapted version of the first part of Messerschmidt *et al.* (2006). More (formal) details of the development of restarting automata can be found in Otto (2006).

An *sRL-automaton* (*simple RL-automaton*) *M* is (in general) a nondeterministic machine with a finite-state control *Q*, a finite characteristic vocabulary *Σ*, and a head with the ability to scan exactly one symbol (word) that works on a flexible tape delimited by the left sentinel ¢ and the right sentinel $.

Let us proceed a bit more formally. A simple *RL-automaton* is a tuple $M = (Q, \Sigma, \delta, q_0, ¢, \$)$, where:

- *Q* is a finite set of states,
- *Σ* is a finite vocabulary (the characteristic vocabulary),
- *¢, $* are sentinels, *{¢, $}* do not belong to *Σ*,
- q_0 from *Q* is the initial state,
- *δ* is the transition relation ≈ a finite set of instructions of the shape (q,a) $\rightarrow_M$ *(p,Op)*, where *q, p* are states from *Q*, *a* is a symbol from *Σ*, and *Op* is an operation, where the particular operations correspond to the particular types of steps (move-right, move-left, delete, accept, reject, and restart step).

For an input sentence $w \in \Sigma^*$, the initial tape inscription is $¢w\$$. To process this input, M starts in its initial state q_0 with its window over the left end of the tape, scanning the left sentinel $¢$. According to its transition relation, M performs *move-right steps* and *move-left steps*, which change the state of M and shift the window one position to the right or to the left, respectively, and *delete steps*, which delete the content of the window, thus shorten the tape, change the state, and shift the window to the right neighbor of the symbol deleted. Of course, neither the left sentinel $¢$ nor the right sentinel $\$$ may be deleted. At the right end of the tape, M either halts and *accepts*, or it halts and *rejects*, or it *restarts*, that is, it places its window over the left end of the tape and reenters the initial state. It is required that before the first restart step and also between any two restart steps, M executes at least one delete operation.

A *configuration* of M is a string $\alpha q \beta$ where $q \in Q$, and either $\alpha = \lambda$ and $\beta \in \{¢\} \cdot \Sigma^* \cdot \{\$\}$ or $\alpha \in \{¢\} \cdot \Sigma^*$ and $\beta \in \Sigma^* \cdot \{\$\}$; here q represents the current state, $\alpha\beta$ is the current content of the tape, and it is understood that the window contains the first symbol of β. A configuration of the form $q_0¢w\$$ is called a *restarting configuration*.

We observe that each computation of an *sRL*-automaton M consists of certain phases. Each part of a computation of M from a restarting configuration to the next restarting configuration is called a *cycle*. The part after the last restart operation is called the *tail*. We use the notation $u \vdash_M^c v$ to denote a cycle of M that begins with the restarting configuration $q_0¢u\$$ and ends with the restarting configuration $q_0¢v\$$; the relation $\vdash_M^{c^*}$ is the reflexive and transitive closure of $\vdash_M^c$.

An input $w \in \Sigma^*$ is *accepted by M*, if there is an accepting computation which starts with the (initial) configuration $q_0¢w\$$. By $L_\Sigma(M)$ we denote the *characteristic language* consisting of all strings accepted by M; we say that M *recognizes (accepts) the language* $L_\Sigma(M)$. By $S_\Sigma(M)$ we denote the *simple language* accepted by M, which consists of all strings that M accepts by computations without a restart step. By *sRL* we denote the class of all sRL-automata.

A *t-sRL-automaton* ($t \geq 1$) is an *sRL*-automaton M which uses at most t delete operations in a cycle and any string of $S_\Sigma(M)$ has no more than t symbols (wordforms).

Remark: The *t-sRL*-automata are two-way automata which allow, in any cycle, to check the whole sentence before reduction (deleting). This reminds us of the behavior of a linguist who can read the whole sentence before choosing the reduction. The automaton should be non-deterministic in general in order to be able to change the order of deleting cycles. That serves for witnessing the independence of some parts of the sentence, see the section about the analysis by

reduction. Another message from this section is that there is a t which creates a boundary for the number of deletions in a cycle and for the size of the accepted irreducible strings.

Based on Messerschmidt *et al.* (2006), we can describe a t-sRL-automaton by *metainstructions* of the form

$(¢ \cdot E_0, a_1, E_1, a_2, E_2, ..., E_{s-1}, a_s, E_s \cdot \$)$, $1 \leq s \leq t$, where

- $E_0, E_1, ..., E_s$ are regular languages (often represented by regular expressions), called the *regular constraints* of this instruction, and
- $a_1, a_2, \ldots, a_s \in \Sigma$ correspond to letters that are deleted by M during one cycle.

In order to execute this metainstruction, M starts from a configuration $q_0¢w\$$; it will get stuck (and so reject), if w does not admit a factorization of the form $w = v_0a_1v_1a_2...v_{s-1}a_sv_s$ such that $v_i \in E_i$ for all $i = 0, ..., s$. On the other hand, if w admits factorizations of this form, then one of them is chosen nondeterministically, and the restarting configuration $q_0¢w\$$ is transformed into $q_0¢v_0v_1\ldots v_{s-1}v_s\$$. To describe also the tails of the accepting computations, we use accepting metainstructions of the form $(¢ \cdot E \cdot \$, Accept)$, where E is a regular language (finite in this case). Moreover, we can require that there is only a single accepting metainstruction for M.

Example: Let us illustrate the power of restarting automata on the formal language L_{Rt}. Let $t \leq 1$, and let $L_{Rt} = \{c_0wc_1wc_2...c_{t-1}w \mid w \in \{a,b\}^*\}$. For this language, a *t-sRL*-automaton M_t with a vocabulary $\Sigma_t = \{c_0,c_1,...,c_{t-1}\} \cup \Sigma_0$, where $\Sigma_0 = \{a,b\}$, can be obtained through the following sequence of metainstructions:

(1) $(¢c_0, a, \Sigma_0^* \cdot c_1, a , \Sigma_0^* \cdot c_2, ..., \Sigma_0^* \cdot c_{t-1}, a, \Sigma_0^* \cdot \$)$,
(2) $(¢c_0, b, \Sigma_0^* \cdot c_1, b , \Sigma_0^* \cdot c_2, ..., \Sigma_0^* \cdot c_{t-1}, b, \Sigma_0^* \cdot \$)$,
(3) $(¢c_0 \ldots c_{t-1}\$, Accept)$.

It follows easily that $L(M_t) = L_{Rt}$ holds.

We emphasize the following property of restarting automata. It plays an important role in our applications of restarting automata.

Definition (Correctness Preserving Property)
A *t-sRL*-automaton M is *(strongly) correctness preserving* if $u \in L_\Sigma (M)$ and $u \vdash_M^{c^*} v$ imply that $v \in L_\Sigma (M)$.

It is rather obvious that all deterministic *t-sRL*-automata are correctness preserving. On the other hand, one can easily construct examples of nondeterministic *t-sRL*-automata that are not correctness preserving.

3.2. The *4-LRL*-automata and related notions

Let us finally introduce the model of automaton proposed for modeling of analysis by reduction for FGD. A *4-LRL-automaton* (*4-level sRL-automaton*) M_{FGD} is a correctness preserving *t-sRL*-automaton. Its characteristic vocabulary Σ is partitioned into four subvocabularies $\Sigma_0, \ldots, \Sigma_3$. M_{FGD} deletes at least one symbol from Σ_0 in each cycle.

Remark: The correctness preserving property of M_{FGD} ensures a good simulation of the linguist performing the analysis by reduction. Similarly as the linguist, the automaton M_{FGD} should not make a mistake during analysis by reduction, otherwise there is something wrong, e.g., the characteristic language is badly proposed. This situation can be fixed by adding some new categories (symbols). The correctness preserving property can be automatically tested. This may be useful for checking and improving a language description in the context of FGD. The request of the deletion of at least one surface wordform in any cycle represents the request of the (generalized) lexicalization of FGD.

Let us inherit the notion $L_\Sigma(M_{FGD})$, characteristic language of M_{FGD}, and $S_\Sigma(M_{FGD})$, the simple language, from the previous subsection. All the notions introduced below are derived from these notions.

As the first step, we introduce an *(analysis by) reduction system* involved by M_{FGD}, and by the set of level alphabets $\Sigma_0, \ldots, \Sigma_3$. It is defined as follows:

$RS(M_{FGD}) = (\Sigma^*, \vdash_{MFGD}^{c}, S_\Sigma(M_{FGD}), \Sigma_0, \ldots, \Sigma_3)$.

The reduction system (by M_{FGD}) formalizes the notion of the analysis by reduction of FGD in an algebraic, non-procedural way. Observe that for each $w \in \Sigma^*$ we have $w \in L_\Sigma(M_{FGD})$ if and only if $w \vdash_{MFGD}^{c*} v$ holds for some string $v \in S_\Sigma(M_{FGD})$.

A *language of level j recognized by* M_{FGD}, where $0 \leq j \leq 3$, is the set of all sentences (strings) that are obtained from $L_\Sigma(M_{FGD})$ by removing all symbols which do not belong to Σ_j. We denote it $L_j(M_{FGD})$. Particularly, $L_0(M_{FGD})$ represents the surface language *LC* defined by M_{FGD}; similarly, $L_3(M_{FGD})$ represents the language of tectogrammatical representations *LM* defined by M_{FGD} (see Section 1).

Now we can define the *characteristic relation* $SH(M_{FGD})$ given by M_{FGD}:

$SH(M_{FGD}) = \{(u,y) \mid u \in L_0(M_{FGD}),\ y \in L_3(M_{FGD})$ and there is a $w \in L_\Sigma(M_{FGD})$ such that u is obtained from w by deleting the symbols not belonging to Σ_0, and y is obtained from w by deleting the symbols not belonging to $\Sigma_3\}$.

Remark: The characteristic relation represents the basic relations in language description, relations of synonymy and ambiguity in language *L*. In other words, it embraces the translation of the surface language *LC* into the tectogrammatical language and vice versa. From this notion, the remaining notions, analysis and synthesis, can be derived.

We introduce the *SH-synthesis by* M_{FGD} for any $y \in LM$ as a set of pairs *(u,y)* belonging to $SH(M_{FGD})$:

$synthesis\text{-}SH(M_{FGD},y) = \{(u,y) \mid (u,y) \in SH(M_{FGD})\}$

The *SH*-synthesis associates a tectogrammatical representation (i.e., string *y* from *LM*) with all its possible surface sentences *u* belonging to *LC*. This notion allows for checking the synonymy and its degree provided by M_{FGD}. The linguistic issue is to decrease the degree of the synonymy by M_{FGD} by the gradual refinement of M_{FGD}.

Finally we introduce the dual notion to the *SH*-synthesis, the *SH-analysis by* M_{FGD} *of* $u \in LC$:

$analysis\text{-}SH(M_{FGD},u) = \{(u,y) \mid (u,y) \in SH(M_{FGD})\}$

The *SH*-analysis returns, to a given surface sentence u, all its possible tectogrammatical representations, i.e., it allows for checking the ambiguity of an individual surface sentence. This notion provides the formal definition for the task of full syntactico-semantic analysis by M_{FGD}.

4. Concluding remarks

The paper presents the basic formal notions that allow for formalizing the notion of analysis by reduction for Functional Generative Description, FGD. We have outlined and exemplified the method of analysis by reduction and its application in processing dependencies and word order in a language with a high degree of free word order. Based on this experience, we have introduced the 4-level reduction system for FGD based on the notion of simple restarting automata. This new formal frame allows us to define formally the characteristic relation for FGD, which renders synonymy and ambiguity in the studied language.

Such a formalization makes it possible to propose a software environment for the further development. It provides a possibility to describe exactly the basic phenomena observed during linguistic research. Further, it allows for studying suitable algorithms for tasks in computational linguistics, namely automatic syntactico-semantic analysis and synthesis.

The presented notions are also useful to show exactly the differences and similarities between the methodological basis of our (computational) linguistic school and the methodological bases of other schools. The basic message given here is to show the possibility of generalizing the principle of lexicalization

trough the layers in order to obtain a checking procedure for FGD via analysis by reduction.

References

Hajič, Jan. 2005. Complex Corpus Annotation: The Prague Dependency Treebank. In *Insight into Slovak and Czech Corpus Linguistics*, ed. M. Šimková, 54–7 3. Veda Bratislava.

Hajičová, Eva, Barbara H, Partee, and Petr Sgall. 1998. *Topic-Focus Articulation, Tripartite Structures, and Semantic Content.* Kluwer, Dordrecht.

Hajičová, Eva. 2006. K některým otázkám závislostní gramatiky. *Slovo a slovesnost* 67:3–26.

Havelka, Jiří. 2005. Projectivity in Totally Ordered Rooted Trees. *The Prague Bulletin of Mathematical Linguistics* 84:13–30.

Holan, Tomáš, Vladislav Kuboň, Karel Oliva, and Martin Plátek. 2000. On Complexity of Word Order. In *Les grammaires de dépendance – Traitement automatique des langues*, ed. S. Kahane,Vol. 41, No. 1, 273–300.

Lopatková, Markéta, Martin Plátek, and Vladislav Kuboň. 2005. Modeling Syntax of Free Word-Order Languages: Dependency Analysis by Reduction. In *Lecture Notes in Computer Science*, Vol. 3658, 140–147.

Messerschmidt Hartmut, František Mráz, Friedrich Otto, and Martin Plátek. 2006. Correctness Preservation and Complexity of Simple RL-Automata. In *Lecture Notes in Computer Science*, Vol. 4094, 162–172.

Mikulová, Marie *et al.* 2006. Annotation on the Tectogrammatical Layer in the Prague Dependency Treebank. Technical report, Prague, MFF UK.

Otto, Friedrich. 2006. Restarting Automata. In: Recent Advances in Formal Languages and Applications. ed. Z Ésik, C. Martin-Vide, V. Mitrana), In *Studies in Computational Intelligence*, Vol. 25, 269–303. Springer, Berlin.

Panevová, Jarmila. 1974. On Verbal Frames in Functional Generative Description. *The Prague Bulletin of Mathematical Linguistics* 22, 3–40.

Petkevič, Vladimír. 1995. A New Formal Specification of Underlying Structure. *Theoretical Linguistics* Vol.21, No.1.

Plátek, Martin: 1982. Composition of Translation with D-trees. In *COLING' 82*, 313–318.

Plátek, Martin, and Petr Sgall. 1978. A Scale of Context-Sensitive Languages: Applications to Natural Language. *Information and Control*. Vol. 38., No 1, 1-20.

Sgall, Petr, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Ed. J. Mey, Dordrecht: Reidel and Prague: Academia.

Sgall Petr, Ladislav Nebeský, Alla Goralčíková, and Eva Hajičová. 1969. *A Functional Approach to Syntax in Generative Description of Language.* New York.

Markéta Lopatková, Martin Plátek, Petr Sgall
Charles University in Prague
Malostranské nám. 25
Prague 1, 118 00, Czech Republic
lopatkova@ufal.mff.cuni.cz

Remark on Author's Share

Building a lexicon, as well as building other comprehensive data resources, goes beyond the potential of one person, which is the reason why it is almost always a collective work of a team of co-authors. This also holds true for the *VALLEX* lexicon. The author of the text presented here is the coordinator of the *VALLEX* project. In close cooperation with Z. Žabokrtský, they developed the structure and the annotation scheme of the lexicon. The author is responsible for application of theoretical framework of FGD and its valency theory to a large amount of linguistic data. She controls the manual processing of verbs (which she also takes part in), including the validation of manual annotations and check for consistency.

The development of NLP systems being necessarily a collective work, this fact is reflected in publication activities as well. The articles in journals, proceedings of top conferences in the field and collective volumes are usually the works of several co-authors with the obvious contribution of each of them.

The author was the principal investigator of the project of the Grant Agency of the Czech Republic called *Valenční slovník českých sloves s komplexní syntakticko-sémantickou informací [Valency Lexicon of Czech Verbs with Complex Syntactico-Semantic Information]*, project No. 405/04/0234. Within this project, the second version of the *VALLEX* lexicon was developed.