Valency Lexicon of Czech Verbs

Zdeněk Žabokrtský

Doctoral Thesis

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University
Prague 2005

Supervisor Mgr. Barbora Vidová-Hladká, Ph.D.
Institute of Formal and Applied Linguistics MFF UK
Malostranské náměstí 25
118 00 Prague 1

Opponents Doc. PhDr. Karel Pala, CSc.
Masaryk University
Fakulty of Informatics
Botanická 68a
602 00 Brno

Doc. RNDr. Karel Oliva, Ph.D.
Czech Language Institute
Academy of Sciences of the Czech Republic
Letenská 4
118 51 Prague 1

This document has been typeset by the author using LaTeX2e with Geert-Jan M. Kruijff's bookufal class and Hans-Peter Kolb's gb4e and cgloss packages.

Abstract

Valency is a property of language units reflecting their combinatorial potential in language utterances. The availability of the information about valency is supposed to be crucial in various Natural Language Processing tasks. In general, valency of language units cannot be automatically predicted, and therefore it has to be stored in a lexicon. The primary goal of the presented work is to create a both human- and machine-readable lexicon capturing valency of the most frequent Czech verbs. For this purpose, valency theory developed within Functional Generative Description (FGD) is used as the theoretical framework.

The thesis consists of three major parts. The first part contains a survey of literature and language resources related to valency in Czech and other languages. Basic properties of as many as eighteen different language resources are mentioned in this part. In the second part, we gather the dispersed linguistic knowledge necessary for building valency lexicons. We demonstrate that if manifestations of valency are to be studied in detail, it is necessary to distinguish two levels of valency. We introduce a new terminology for describing such manifestations in dependency trees; special attention is paid to coordination structures. We also preliminarily propose the alternation-based lexicon model, which is novel in the context of FGD and the main goal of which is to reduce the lexicon redundancy. The third part of the thesis deals with the newly created valency lexicon of the most frequent Czech verbs. The lexicon is called VALLEX and its latest version contains around 1600 verb lexemes (corresponding to roughly 1800 morphological lemmas); valency frames of around 4400 lexical units (corresponding to the individual senses of the lexemes) are stored in the lexicon. The main software components of the dictionary production system developed for VALLEX are outlined, and selected quantitative properties of the current version of the lexicon are discussed.

Acknowledgements

At the very beginning, I would like to thank my colleagues at IFAL for creating a friendly and stimulating atmosphere. In particular, I owe special thanks to Markéta Lopatková, first for her patience with me during our countless discussions, and second for helping my restless mind to hold the research course. Without her, this thesis would hardly have been written and I would have been just rambling from one topic to another all the time.

I am also grateful to Karolína Skwarska and Václava Benešová for contributing to VALLEX. Without their careful lexicographic work, VALLEX may never have seen the light of day, definitely not in its present shape and size.

I would like to thank those who created the great adventure of the Prague Dependency Treebank project, which I had the luck to participate in (even if only for a while). It was a very special experience with a very special team, and I appreciate it a lot.

I am also obliged to Petr Pajas, for always being here to listen and help. Thanks to him, I learned again what is the joy of programming.

My deep gratitude is due to those who provided me with numerous comments on the draft of this thesis, especially Jarmila Panevová, Petr Sgall, my supervisor Barbora Vidová-Hladká, Markéta Lopatková, and Silvie Cinková.

Finally, I would like to thank my Mum and Dad for all their love and support, for letting me play, for enduring my forestry experiments, and also for ... but I really wonder where their life energy comes from, because as it seems to me now, they will be soon younger than I am and I will be the elderly one ... and of course I cannot forget my sweet little sister with her warm and loving 'Hey, the stranger from Prague is back again!'

Contents

Chapter 1

Introduction

> Valency—the range of syntactic elements either required or specifically permitted by a verb or other lexical unit.
> **Concise Oxford Dictionary of Linguistics**

1.1 The term "valency"

The word "valency" comes from the Latin "valentia" (strength) or "valere" (to be strong).[1,2] In 1852 Edward Frankland announced his theory of valency, that is, each atom has a certain "valency," or capacity for combining with a definite number of other atoms. One hundred years later, Lucien Tesnière introduced this term into linguistics and used it in the context of syntactic analysis of a sentence ([Tesnière, 1959]). When analysing the sentence, he started with the verb and classified its "subordonnés immédiats" into "actants" and "circonstants". He saw a resemblance between the chemical valency and the ability of verbs to bind at most a certain number of actants, and called the latter valency too.

Active valency (what arguments a language unit requires) and passive valency (to what other language units it can attach) are occasionally distinguished in literature (recently e.g. in [Nasr and Rambow, 2004]). In this thesis, the term of valency will be used exclusively in the sense of active valency.

As we will see, nowadays there is a whole bunch of various models of valency, each having its own pecularities, usually not justifiable by the specifics of the studied language. Why is that?

The motto of this chapter is vague enough to match all the approaches to valency which we are aware of. Though it provides us only with a very inexplicit feeling about the term in question, it excellently explains the undesirable diversity mentioned above by alluding the main sources of this diversity:

[1] http://www.wordreference.com/definition/valency.htm

[2] Although we cannot guarantee the stability of the URL pointers on the Internet, we use them througout this thesis. It has three main reasons: Either we were not able to find any proper reference to a "printed" source of the information, or such source even does not exist because of its nature (e.g. on-line databases), or we simply believe that the reader might find the URL useful.

- Does the term "range" stand for the sole number of the elements, or a set of them, or a sequence of them, or a structured hierarchy of them?
- "Syntactic" – which syntax: constituency or dependency (both having dozens of variants nowadays)? Or a combination of both? And is that surface syntax or deep syntax, or do we need both?
- What properties are to be specified for the "elements"?
- "Required" – does that strictly mean that such element must be uttered in the sentence, or that it must be at least known to the speaker (or to the listener?), or something completely else?
- "Permitted" in which sense? By what (or by whom)? And if things can be "specifically permitted", can they be also "specifically prohibited"?
- "Verb" – does it cover also the auxiliary verbs (and light verbs, support verbs, modal verbs ...). Do they have their own valency?
- "Lexical unit" – is that a word, or a word form, or even a multi-word expression, or a word form used in a given sense, or a word form given the sense and the context, or ...?

In this thesis, we use the term valency in a relatively wide sense, similarly e.g. to [Briscoe, 2001]:

> I use the term valency in an extended sense as a relatively theory-neutral term to refer to lexical information concerning a predicate's realization as a single or multiword expression (such as a phrasal verb), the number and type of arguments that a particular predicate requires, and the mapping from these syntactic arguments to a semantic representation of predicate-argument structure which also encodes the semantic selectional preferences on these arguments. Thus, I use the term valency (frame) to subsume (syntactic) subcategorization and realization, argument structure, selectional preferences on arguments, and linking and/or mapping rules which relate the syntactic and semantic levels of representation.

Besides the term "valency", also other terms denoting the same (or a similar) concept are used:

- "Valency" and "subcategorization" are often treated as synonymous (or the reader is left confused whether they are synonymous or not), e.g. in [Fujita and Bond, 2002], but there are strictly distinguished elsewhere.

- In some other studies, the term "argument structure" is used for denoting the property of a lexical unit, and the term "valency" stands only for the sole number of arguments in the argument structure.
- Following [Pauliny, 1943], the term "intention" is ocasionally used by some Czech and Slovak authors, either istead of valency, or instead of deep valency.
- Counterparts of surface and deep are by some authors presented as a part of a single and more complex representation, called "goverment pattern" ([Mel'čuk and Zholkovsky, 1984]), "complex sentence pattern" ([Daneš, 1994]), "head's argument structure" ...
- Other terms such as "stereotypical syntagmatic patterns" (introduced in [Pustejovsky et al., 2004]), "government" (not in the sense of Chomskyan Government and Binding) or "case assignment" are ocasionally used; their relation to the term "valency" seems to be obvious but is not explicitely stated.

1.2 Motivation and goals

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

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

  (1) Stali se matematiky.
  They became mathematicians.

  (2) Báli se matematiky.
  They worried 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* (sentence 1) does not contain[3] acc. in its valency frame, and no frame of the verb *bát se* (sentence 2) contains acc.[4]

[3] In this context, we say 'frame X contains Y' for expressing the fact that some slot of the valency frame X is prototypically realized by the form (direct or preposition case, etc.) Y on the surface.

[4] The possibility of nominative is excluded in both sentences according to subject-verb agreement.

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

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

  Noun phrase *jeho bratra* (preceded by no preposition) can be gen.sg. or acc.sg. Verb *ptát se* ('to ask') allows only the former possibility.

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

  (4) Nechala ho spát.
  She let him sleep.

  (5) Začala ho milovat.
  She started to love him.

  In sentence (4) the pronoun *ho* ('him') can depend only on the preceding verb (since valency frame of *spát* ('to sleep') does not contain accusative), whereas 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[5] 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 description.

  Different meanings of the same word are often indicated by a change in the valency frames.

- *'Semantic analysis'.* Examples:

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

[5] Similar claim holds for phrase structure of the given sentences.

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

Prepositional groups most frequently represent adverbials (like in sentence 9), however, they can also stand for verbal arguments (like in 10), which is a crucial difference in most semantically or logically motivated approaches. The role of the prepositional group *po Petrovi* 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 to machine translation (MT). Since the existence of a valency dictionary would lead to a higher quality of the respective submodules of such MT system, it should also increase the quality of the resulting translation.

As it was illustrated above, valency lexicon is an important information resource for NLP applications. However, when the present author started working in the field of computational linguistics, there was no publicly available high-quality machine-readable (and also machine-tractable) lexicon of Czech verbs. Thus the primary goal of the presented work was to build such a lexicon and make it available to other researchers.

The primary goal implied secondary goals. First, it was necessary to collect linguistic knowledge which comes into play when building the lexicon. VALLEX is based on the formal linguistic framework called Functional Generative Description (FGD); however, terminology and techniques inspired by other frameworks will be used in this thesis too (certain borrowings will be necessary because FGD has not been focused on practical lexicography). Second, a dictionary production system had to be developed: it comprises a system of software tools for manual annotation, tools for searching, tools for format conversions, specification of the data formats, annotation methodology, work-load distribution etc.

1.3 Structure of the thesis

The rest of the presented thesis consists of three major parts:

- The first part contains a survey of literature and language resources related to valency in Czech (Chapter 2) and other languages (Chapter 3). As many as eighteen different language resources are mentioned in this part.
- In the second part (formed by Chapters 4 and 5), we try to gather (and add to) the dispersed linguistic knowledge necessary for building a valency lexicon. In Chapter 4 we study how the lexicon should be structured (in terms of lexemes and lexical units), while in Chapter 5

we turn to the manifestation of valency in deep and surface dependency syntactic structures.

- In Chapter 5 we describe the technical details related to the process of the creation of VALLEX. All main components of the dictionary production system are mentioned and some empirical properties of the created lexicon are discussed.

Finally, Chapter 7 summarizes the contribution of this thesis.

Chapter 2

Studies on Valency in Czech

> Most importantly, most of the linguistic properties that must be considered for text processing are not emergent properties of the texts at all but crucially depend on *l'arbitraire du signe*, the arbitrary relation between a symbol and what it symbolizes.
>
> **Martin Kay**

This chapter has two parts. In the first part, we shortly describe two theoretical frameworks that have been dominating the studies on valency in the Czech linguistic literature for the last three decades, namely the Sentence-pattern Model (Section 2.1.1) and the valency theory developed within Functional Generative Description (Section 2.1.2). In the second part, we briefly present five lexical resources (one printed and four electronic) related to the valency of Czech verbs (Sections 2.2.1 to 2.2.5).

2.1 Theoretical Approaches

2.1.1 The Sentence-Pattern Model

Theoretical essentials of the Sentence-Pattern Model (SPM) were formulated in mid-fifties. The results of the following long-term collective research headed by František Daneš were summarized [Daneš and Hlavsa, 1987] (see also [Daneš, 1994] for references). In SPM, the three-layer approach to sentence analysis was developed: (a) the level of the grammatical sentence structure, (b) the level of the semantic sentence structure, (c) the level of the communicative organization of the utterance.

In the early years, first the concept of *grammatical sentence patterns* (GSP) was specified. Subsequently, *semantic sentence patterns* (SPP) were suggested, and the correlation between SSP and GSP called *complex sentence pattern* (CSP) was integrated into the language description. The GSP and SSP instantiated in the sentence "The farmed killed a duckling" look as follows:

GSP: S_{nom}	-	VF	-	S_{acc}
SSP: agent	-	act. caus.	-	patient

With SPM, also the distinction between constitutive (either obligatory or potential) and non-constitutive (facultative) sentence components was elaborated.

Verb predicates, as the central components of sentence structure, received most attention within SPM. Around 2000 verbal lexemes of Czech were preliminarily examined, and around 300 most frequent ones were analyzed in detail in [Daneš and Hlavsa, 1987]. Later, SPM was used in lexicographic praxis when creating [Svozilová et al., 1997] (see Section 2.2.3).

One of the important contributions of SPM is also the system of verb classification; especially the following classes are elaborated: verbs of motion, verbs of manipulation, verbs of speaking, thinking and perception, verbs of change, and verbs of elementary processes.

SPM is used as a starting point also in [Karlík, 2000], where the so called Modified valency theory is presented. In this work, the Sentence-Pattern Model is merged with elements from the Western generative stream (for instance, external and internal valency positions are distinguished).

2.1.2 Valency theory in Functional Generative Description

Functional Generative Description (abbrev. FGD) is a system of the description of natural language developed in the Prague group of mathematical linguistics since mid-sixties ([Sgall, 1967]). FGD is a (dependency-based) stratificational approach, i.e. it decomposes the description of language into a system of levels (strata). Five levels of representation were proposed in the original version:

- tectogrammatical representation
- surface-syntactic representation[1]
- morphological representation
- morphonological representation
- phonetic representation[2]

Valency theory, elaborated in [Panevová, 1974] and [Panevová, 1980], is one of the core components of FGD, especially of its tectogrammatical level. The theory refers back to [Tesnière, 1959] (and was inspired also by [Helbig and Schenkel, 1969]) and is based on the following postulates (we will mention them only very briefly, as their more detailed description is available in many recent publications, e.g. [Skoumalová, 2001] [Lopatková, 2003] or [Urešová, 2004]):

- verbal complementations (dependents) can be classified either as *inner*

[1] In the recent versions of FGD (roughly since 1990's), surface syntax is not treated as an autonomous level of language description any more.

[2] As in other contemporary symbolic stratificational approaches, phonetics received much less attention when compared to higher levels. In our opinion, this was partly due to the fact that as the time went by, purely probabilistic models rather than linguistically motivated solutions were relatively successful in the field of spoken language processing.

participants (*actants*,*arguments*)[3] or as *free modifications* (*adjuncts*),

- the relation between the governor and its dependent is labeled with a functor; five functors for actants are distinguished: actor, patient, addressee, origin, effect; functors also distinguish between various types of temporal, locational, causative and other free modifications,
- both actants and free modifiers can be either obligatory, or optional for the given verb; the so called *dialogue test* was introduced as a criterion for distinguishing obligatory and optional dependets ([Panevová, 1974]),
- a valency frame (in the narrow sense) contains only actants and those free modifiers which are obligatory for the given verb,
- a verb's valency in the wider sense concerns also all of its optional adjuncts; the present thesis is not concerned with this aspect,
- the concept of *shifting of cognitive roles* is used when assigning functors to an actant: if a verb has one actant, it is always actor; if there are two, one is always actor and the other is patient, no matter what its cognitive role with respect to the verb is; only if there are three or more actants, semantic criteria come into play.

2.2 Language Resources

2.2.1 Valency lexicon BRIEF

Electronic valency lexicon BRIEF ([Pala and Ševeček, 1997]) created at the Masaryk University in Brno was compiled from several existing printed dictionaries for Czech, especially [SSJČ, 1978]. It contains around 15,000 verbs.

Sample from the lexicon is given in Figure 2.1. The format of the lexicon is not easily readable, therefore we will try to describe it at least very briefly here (a complete description can be found in [Horák, 1998]). For each verb, the lexicon contains a list of frames separated by comma. Frame is a sequence of elements separated with dash, where each element is represented as a sequence of attribute-value pairs. Attributes are denoted with lower case letters, and values are denoted either as capital letters, or they are delimited by braces. The following attributes are used: h – semantic feature (with values T for thing and P for person), c – morphological case, r – preposition, s – infinitive or subordinating clause, e – negated subordinating clause, i – idiomatic expression, v – other features, z – comment.

The disadvantage of this lexicon is that if an argument of a verb in a given sense can be expressed in more ways, then they have to be captured in different frames and thus the number of frames is in some cases inadequately high (experiments with merging such frames are presented in

[3] Although the word actant comes from French and is still not contained in most English dictionaries, many authors use it routinely in English texts related to valency (e.g. [Mel'čuk, 2004]) and so do we.

```
běžet <v>hTc2r{z},hTc2r{do},hTc4r{na},hPTc3r{ke},hPTc4r{pro},hPTc7r{za},
hTc2r{kolem},hPc3r{proti},hPc3r{proti},hTc4r{o},hTc4z{motor pravidelně}
bičovat <v>hPTc4
bídačit <v>hPc4,hPc4-hTc7
bídačit se <v>
biflovat <v>hTc4,hTc2r{z},hTc4r{na},hTc4-hTc2r{z},hTc4-hTc4r{na}
biflovat se <v>hTc4,hTc2r{z},hTc4r{na},hTc4-hTc2r{z},hTc4-hTc4r{na}
bilancovat <v>hTc4
bílit <v>hTc4,hTc4-hTc7
bířmovat <v>hPc4
bít <v>hPTc4,hPTc4-hTc7,hPc6r{po},hTc4r{na},hTc7-hTc4r{o},hTc7-hTc2r{do}
bít se <v>hPTc7r{s},hPTc7r{s}-hTc3r{kvůli},hPc7r{s},hTc4r{o},
hPc7r{s}-hTc4r{o},hPTc4r{za}
blábolit <v>hTc4,hTc6r{o},hTc4-hTc6r{o}
```

Figure 2.1: Sample from Valency lexicon BRIEF.

[Skoumalová, 2001]). As the lexicon contains no example usages or other similar clues, it is sometimes difficult for the user to judge which frames correspond to which senses.

2.2.2 Czech syntactic lexicon

Czech syntactic lexicon ([Skoumalová, 2001], [Skoumalová, 2002]) was created by an automatic conversion from the BRIEF lexicon and thus contains the same amount of verbs (15,000). During the conversion, the lexicon was slightly restructured (some frames were merged) and enriched with new linguistically relevant information (functors, diatheses, reflexivity and control) using an algorithm based on linguistic observations. The newly created lexicon uses the theoretical background of Functional Generative Description.

Sample entry for the verb *brzdit* (to brake, to inhibit):

```
brzdit R--s[i1]1(hPTc1)2[hPTc4]%$
```

Explanation:

`R`	active voice
`--`	empty position for reflexive particle
`s[i1]`	inherent subject in position marked with 1
`1, 2`	actants (1 – Actor, 2 – Patiens)
`h`	semantic feature (P – person, T – thing)
`c`	morphological case
`%`	possibility of passive
`$`	possibility of reflexive passive

This lexicon inherits some disadvantages of BRIEF. For instance, if a given verb has two different senses with the same frames, then these senses remain undifferentiated in the lexicon.

2.2.3 Valency lexicon "Slovesa pro praxi"

The valency lexicon *Slovesa pro praxi* ([Svozilová et al., 1997]) contains a detailed analysis of 767 most frequent Czech verbs. The sample from the lexicon is depicted in Figure 2.2. The lexicon is based on the theoretical framework of sentence pattern introduced in [Daneš and Hlavsa, 1987]. In the foreword, the authors describe the content of the lexicon entries as follows:

> Valency potential of each sentence predicate has in our lexicon three (mutually related) levels of presentation: general pattern, explicit pattern, and morphologico-syntactic analysis of occupation of individual valency positions, to which also a semantic component is added: information about selective semantic features of tendencies, which limit the choice of concrete expressions occupying the valency positions. All items are then documented in the example part.

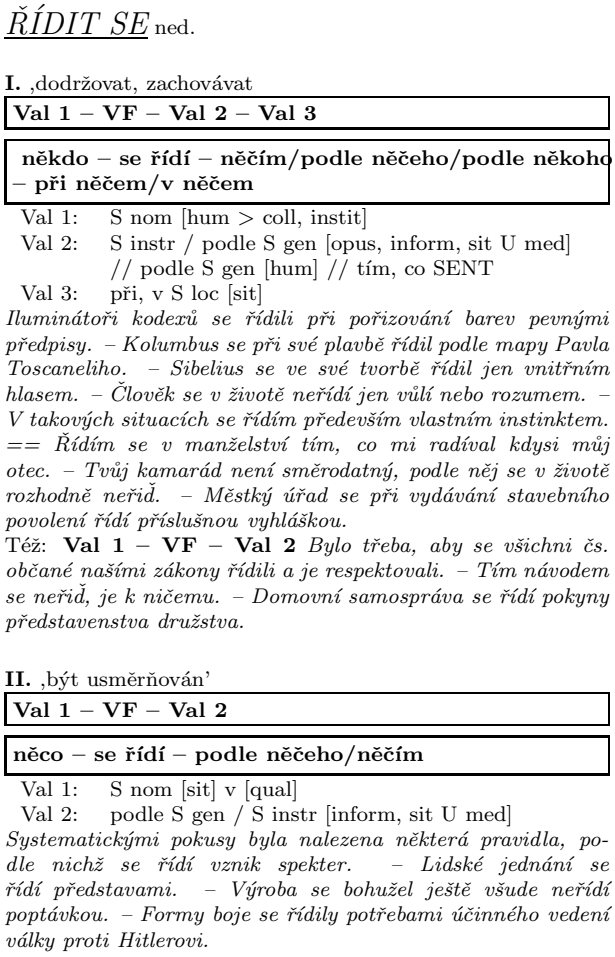
The authors distinguish around 50 quite diverse selective semantic features (and also some of their combinations are acceptable). We present only a few of them:

- anim – animal beings
- circ – circumstantial
- coll – collectivity
- fin – finality
- med – medium
- mod – modus
- opus – artifact
- orig – origin
- plant – plat
- reciproc – reciprocity
- signum – sign
- totum – totality

2.2.4 PDT-VALLEX

PDT-VALLEX ([Hajič et al., 2003], [Urešová, 2004]) is a valency dictionary gradually created and used during the annotation of the Prague Dependency Treebank, version 2.0.[4] The valency theory of FGD is used as the theoretical

[4]To appear at http://ufal.mff.cuni.cz/pdt2.0/. PDT-VALLEX data will be distributed together with PDT 2.0.



ŘÍDIT SE ned.

I. ‚dodržovat, zachovávat

Val 1 – VF – Val 2 – Val 3

někdo – se řídí – něčím/podle něčeho/podle někoho – při něčem/v něčem

Val 1: S nom [hum > coll, instit]
Val 2: S instr / podle S gen [opus, inform, sit U med] // podle S gen [hum] // tím, co SENT
Val 3: při, v S loc [sit]

Iluminátoři kodexů se řídili při pořizování barev pevnými předpisy. – Kolumbus se při své plavbě řídil podle mapy Pavla Toscaneliho. – Sibelius se ve své tvorbě řídil jen vnitřním hlasem. – Člověk se v životě neřídí jen vůlí nebo rozumem. – V takových situacích se řídím především vlastním instinktem. == Řídím se v manželství tím, co mi radíval kdysi můj otec. – Tvůj kamarád není směrodatný, podle něj se v životě rozhodně neřiď. – Městký úřad se při vydávání stavebního povolení řídí příslušnou vyhláškou.

Též: **Val 1 – VF – Val 2** *Bylo třeba, aby se všichni čs. občané našími zákony řídili a je respektovali. – Tím návodem se neřiď, je k ničemu. – Domovní samospráva se řídí pokyny představenstva družstva.*

II. ‚být usměrňován'

Val 1 – VF – Val 2

něco – se řídí – podle něčeho/něčím

Val 1: S nom [sit] v [qual]
Val 2: podle S gen / S instr [inform, sit U med]

Systematickými pokusy byla nalezena některá pravidla, podle nichž se řídí vznik spekter. – Lidské jednání se řídí představami. – Výroba se bohužel ještě všude neřídí poptávkou. – Formy boje se řídily potřebami účinného vedení války proti Hitlerovi.

Figure 2.2: Sample from the lexicon "Slovesa pro praxi".

framework. Entries of PDT-VALLEX contain individual senses of verbs and certain verbal nouns and adjectives that have been found in the treebank texts annotated at the tectogrammatical layer. Each sense contains a valency frame with semantic, syntactic and morphological information about its semantically obligatory and/or optional dependents. There are around 5500 verbs, 3700 nouns and 800 adjectives in the PDT-VALLEX.

Frame instances occurring in the treebank are interlinked with their dictionary entries, which made it possible to check whether the frame instances exactly match the frame specification in the lexicon.

Sample entry from the PDT-VALLEX is depicted in Figure 2.3.

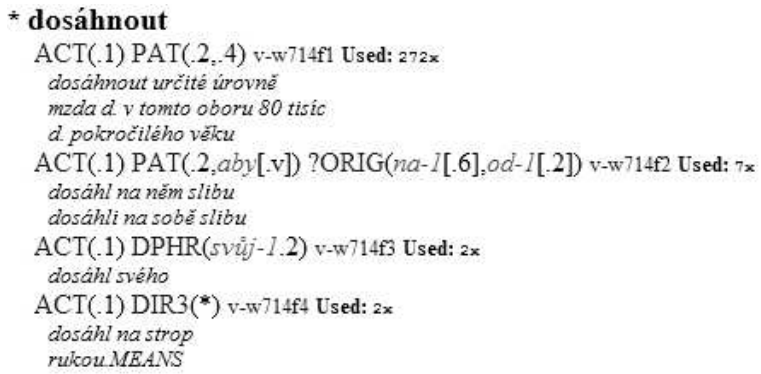


Figure 2.3: Sample entry from PDT-VALLEX.

2.2.5 VerbaLex

VerbaLex[5] ([Hlaváčková and Horák, 2005]), originally called FIMU VALLEX, is a recently developed lexical database that enriches the Czech part of EuroWordNet data ([Vossen, 1998]) with valency frames,[6] using adapted data formats and some of the tools originally developed for VALLEX.

The lexical units in EuroWordNet are organized into synsets (sets of synonyms). Entries in VerbaLex contain lemmata with synonymic relation and with common valency frame. The authors claim that the main difference between VALLEX and VerbaLex valency frames is that the latter uses a two-level system of semantic roles derived from Princeton WordNet Base Concepts. For instance, VerbaLex distinguishes whether the actor of the verb should be a person or an animal. There are more than 3200 verbs in around 1650 VerbaLex entries.

A sample of the VerbaLex entry is depicted in Figure 2.4.

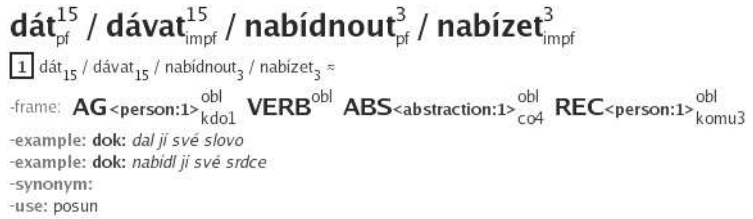


Figure 2.4: Sample entry from VerbaLex (synset with the meaning *to give/to offer*).

[5] http://nlp.fi.muni.cz/verbalex/

[6] The need of adding valency patterns into WordNet structures is discussed also in other languages, e.g. in German ([Kunze and Rösner, 2004]) or Spanish ([Civit et al., 2005]).

Chapter 3

Studies on Valency in Other Languages

> Unfortunately you don't have the thousand years
> and the thousand people.
> **Eduard Hovy**

There are tens of different theoretical approaches, tens of language resources and hundreds of publications related to the study of valency in various natural languages. Some of them crystallized after several decades of linguistic research, but many others describe just a small experiment performed a couple of days before a conference deadline. Some of them have resulted in an extensive language resource, be it a printed dictionary (or a book appendix) or an electronic database, but many others present just an isolated phenomenon without sufficient empirical evidence. It goes beyond the scope of this thesis (and probably beyond the capability of a human individual) to give an exhaustive survey of all these enterprises.

In any case, it is surprising that most of them have remained isolated or even generally unknown (mostly due to extralinguistic reasons), and that the general NLP community is probably aware only of two or three most prominent projects from this field (be it called valency, predicate-argument structure, or frame semantics). In the following sections, we try to present a little bit wider outline of works related to valency. Only the basic properties are mentioned, but we try to provide the reader with authentic samples from the individual resources. We select only those works that have resulted in a language resource (either electronic or printed) containing at least several hundred verbs.

3.1 Dictionary of Valency and Distribution of German Verbs

One of the most remarkable attempts at formal description of valency (especially when considering the time of the creation) is [Helbig and Schenkel, 1969]. It contains around 350 German verbs, each of them described in three steps:

I. In the first step, the number of "verb partners" (*Anzahl der Mitspieler*) is determined, e.g.: *erwarten*$_2$ (*to expect*), *rauben*$_{2(3)}$ (*to rob*); the parentheses contain the number of optional partners.

II. In the second step, the forms of the "syntactic surrounding" are described. Abbreaviations such as Sn (substantive in nominative), Sa, Sd, Sg, NS (*Nebensatz*, subordinating clause), I (infinitive without *zu*), Inf (infinitive with *zu*) etc. If the partner is optional, then it is written in parentheses. If there are more alternative forms, then they are separated by a slash.

III. In the third step, the "semantic surrounding" is described in terms of semantic features of individual partners, such as Hum (human), +Anim (animate), -Anim (inanimate), Abstr (abstract), Loc (locational), Temp (temporal) etc.

A sample containing the verb *hören* is presented in Figure 3.1.

hören

I.	hören$_{2(3)}$ (V1 = wahrnehmen, aufnehmen)	
II.	hören → Sn, Sa/NS$_{dass,ob,w}$, (I)	
III.	Sn →	+ Anim (Das Kind, die Katze hört den Fremden.)
	Sa →	1. + Anim (Er hört das Kind, den Hund.)
		2. Abstr (Er hört Musik)
		3. Act (Er hört das Brüllen.)
	NS →	Act (Er hört, dass er kommt / ob er kommt / wer kommt.)
	I →	Act (Er hört sie kommen.)

I.	hören$_{2}$ (V2 = gehorchen, reagieren)	
II.	hören → Sn, pS	
III.	Sn →	1. + Anim (Das Kind, der Hund hört auf ihn.)
		2. Abstr (als Hum) (Die Betriebsleitung hö rt auf seinen Rat.)
	p = auf,	
	pSa →	1. hum (Die Schüler hören auf den Lehrer.)
		2. Abstr (Die Schüler hören auf seine Worte.)

Anmerkungen:
1. Bei V1 ist I als 3.Mitspieler nur möglich, wenn als 2.Mitspieler Sa (nich wenn NS) erscheint. I zusammen mit Sa kann als Ersatz für NS angesehen werden ("Ich höre ihn kommen" – "Ich höre, dass er kommt").
2. Vereinzelt ist bei V 1 für Sa auch – Anim, möglich, aber nur bei sich bewegenden Objekten ("Er hört das Flugzeug". Aber: "*Er hört den schrank").

Figure 3.1: Description of valency of the verb *hören* (to hear) in [Helbig and Schenkel, 1969] (arranged sample).

3.2 Syntactic Generative Dictionary of Polish Verbs

Five volumes of the valency lexicon of Polish verbs ([Polański, 1992]) were issued from 1980 to 1992. The authors used the word "generative" (contained in the title of the lexicon) to refer to two postulates of generative linguistics: (1) explicitness of language description, (2) emphasis on creative nature of language.

The lexicon entry contains the following information:

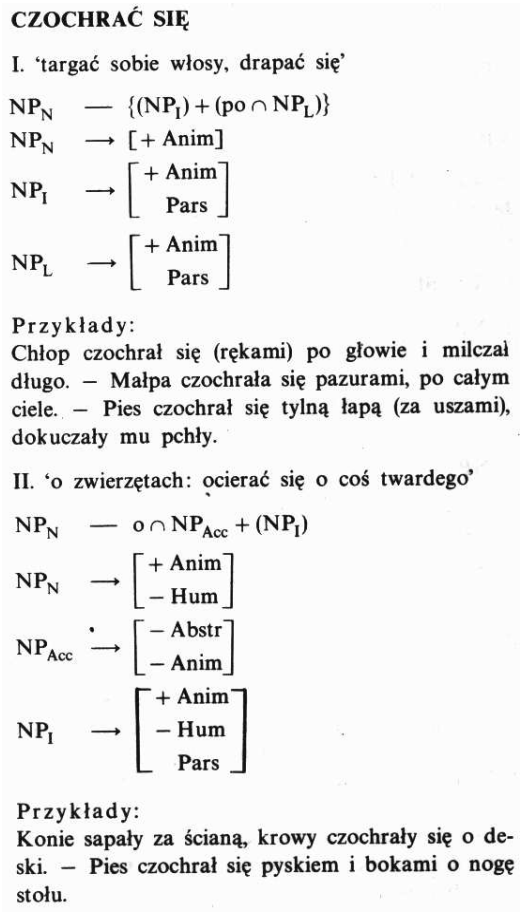
- infinitive of the headword
- division to sub-entries according to the individual senses
- sentence scheme composed of symbols for the individual parts of the sentence structure, e.g. NP (Noun Phrase) with further symbols in subscript for form specification (e.g. NP_{Acc} case). One of the senses of the verbs *cmokać – cmoknać*:
  NP_N – $\{(\mathrm{NP}_I) + (\mathrm{na} \bigcap \mathrm{NP}_{Acc}) + (\mathrm{NP}_{Caus})\}$
  Besides NP, also other symbols are used, e.g. OR (direct speech), $K \bigcap S$ (interrogative particle or pronoun and clause).
- semantic characteristics of the nominal parts (semantic features such as [+/-Anim] [+/-Abstr]; if the features are to be applied together, then they are written below each other, otherwise they are alternative and are written one after another.
- information about the possibility of forming passive (unless the existence or non-existence of passive of a given verb is implied by a general rule)
- examples of usage
- phraseologisms

A lexicon sample is depicted in Figure 3.2.

3.3 Valency dictionary of Slovak verbs

The valency dictionary of Slovak verbs [Nižníková and Sokolová, 1998] contains grammatical and semantic characteristics of 625 most frequent Slovak verbs. Entries are structured as follows:

- infinitive of the entry verb
- list of all its lexias (senses)
- for each lexia:
  - description of the meaning of the lexia
  - semantic structure
  - synonyms



CZOCHRAĆ SIĘ

I. 'targać sobie włosy, drapać się'

NP_N — $\{(NP_I) + (po \cap NP_L)\}$

$NP_N \longrightarrow [+ Anim]$

$NP_I \longrightarrow \begin{bmatrix} + Anim \\ Pars \end{bmatrix}$

$NP_L \longrightarrow \begin{bmatrix} + Anim \\ Pars \end{bmatrix}$

Przykłady:
Chłop czochrał się (rękami) po głowie i milczał długo. – Małpa czochrała się pazurami, po całym ciele. – Pies czochrał się tylną łapą (za uszami), dokuczały mu pchły.

II. 'o zwierzętach: ocierać się o coś twardego'

NP_N — $o \cap NP_{Acc} + (NP_I)$

$NP_N \longrightarrow \begin{bmatrix} + Anim \\ - Hum \end{bmatrix}$

$NP_{Acc} \longrightarrow \begin{bmatrix} - Abstr \\ - Anim \end{bmatrix}$

$NP_I \longrightarrow \begin{bmatrix} + Anim \\ - Hum \\ Pars \end{bmatrix}$

Przykłady:
Konie sapały za ścianą, krowy czochrały się o deski. – Pies czochrał się pyskiem i bokami o nogę stołu.

Figure 3.2: Sample from [Polański, 1992] with the verb *czochrać sie* (to scratch oneself).

- valency structure
- participant characteristics
- examples of usages
- the possibility of transformed (derived) structures

The Sample of a lexicon entry is reproduced in Figure 3.3.

3.4 FrameNet

The Berkeley FrameNet project[1] ([Fillmore, 2002], [Fillmore et al., 2002]) is aimed at creating an on-line lexical resource for English, based on frame semantics ([Fillmore, 1968]). Its goal is to document the range of semantic

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

RÁSŤ ndk

rásť 1	rastom sa zväčšovať, vyvíjať sa
rásť 2	vyrastať, dospievať
rásť 3	vyskytovať sa, rodiť sa niekde
rásť 4	vznikať, vzmáhať sa, zveľaďovať sa
rásť 5	zväščovať svoj objem, intenzitu, význam
rásť 6	zdokonal'ovať sa vo vývine
rásť 7	vznikať, vytvárať sa, vyvíjať sa

rásť 1 rastom sa zväščovať, vyvíjať sa
SŠ: -A, +D, +R, +M
SYN: zväčšovať sa, vyvíjať sa
VŠ: /Sn/ – VF
Sn: živý organizmus [ANIM/PLANT]: PROCnd
Ako tie deti rastú!
Pozri, už rastie tráva.
To dieťa rastie ako z vody.
TRANSF: 0

rásť 2 vyrastať, dospievať
SŠ: -A, +D, +R, +M
SYN: vyrastať, dospievať
VŠ: /Sn/ – VF – ADVloc
Sn: osoba, ktorá vyrastá: PROCnst:
ADVloc/mod: miesto, spôsob
Rástol som na dedine.
Naše deti rástli v hojnosti.
TRANSF: 0

rásť 3 vyskytovať sa, rodiť sa niekde
SŠ: -A, -D, +R, -M
SYN: vyskytovať sa, rodiť sa niekde
VŠ: /Sn/ – VF – ADVloc
Sn: ten/to, kto/čo niekde rastie [ANIM/PLANT]: STATnst
ADVloc: miesto
V lese rastú huby.
Kde rastú pekné dievčatá?
Niektoré liečivé rastliny rastů vo vol'nej prírode.
TRANSF: 0

rásť 4 vznikať, vzmáhať sa, zvel'aďovať sa
SŠ: -A, +D, +R, +M
SYN: rásť 4 vznikať, vzmáhať sa, zvel'aďovať sa
VŠ: /Sn/ – VF – (ADVloc/mod)
Sn: to, čo vzniká [KONKR/REG]: PROCnst
ADVloc/mod: miest, spôsob
V meste rastú nové štvrte.
Rastú nové firmy a podniky.
TRANSF: 0

rásť 5 zväščovať svoj objem, intenzitu, význam
SŠ: -A, +D, +R, +M
SYN: zväščovať sa, narastať
VŠ: /Sn/ – VF
Sn: to, čo rastie [QUAL/SIT/MENT]: PROCnst
Jeho vplyv stále rástol.
Rastú rady nezamestnaných.
Výroba rastie iba pomaly.
TRANF: 0

rásť 6 zdokonal'ovať sa vo vývine
SŠ: -A, +D, +R, +M
SYN: zdokonal'ovať sa, vyvíjať sa
VŠ: /Sn/ – VF – ADVmod/asp
Sn: ten, kdo rastie [HUM]: PROCnst
ADV mod/asp: spôsob, aspekt, oblasť
Spisovatel' rástol l'udsky aj umelecky.
Človek rastie prekonávaním prekážok.
TRANSF: 0

rásť 7 vznikať, vytvárať sa, vyvíjať sa
SŠ: -A, +D, +R, +M
SYN: zdokonal'ovať sa, vyvíjať sa
VŠ: /Sn/ – VF – z Sg
Sn: to, čo vzniká [ANIM/QUAL]¿ PROCnst
z Sg: to, z čoho Sn rastie [ANIM/ORIG]: OBJ
Rastlinky rastú zo semienka.
Z milého šteňaťa rástol nebezpečný pes.
Rastie z neho zlodej.
Rastie z teba pekný kvietok!
TRANSF:0

Figure 3.3: Sample from the Valency lexicon of Slovak verbs ([Nižníková and Sokolová, 1998]) with the verb *rásť* (to grow).

Causation			
Cause		Affected	Effect
The wind	*caused*	*the tree*	*to sway.*

Communication					
Speaker		Message	Addressee	Topic	Medium
Pat	*communicate*		*with Kim*	*about the festival.*	
Pat	*communicate*		*with Kim*		*by letter.*
Pat	*communicate*	*the message*	*to me.*		

Reciprocality		
Protagonist-1		Protagonist-2
Pat	*fought*	*with Kim.*
Pat and Kim	*fought.*	

Figure 3.4: FrameNet: Examples of semantic frames.

and syntactic combinatoric possibilities of each word (especially verbs and 'frame-bearing' nouns) in each of its senses.

Semantic frames in FrameNet are representations for prototypical situations or states, and lexical units are grouped into such frames. Each frame provides its set of semantic roles (which are thus specific for the given group of lexical units), such as Speaker, Message or Topic. Basic semantic frames 'Causation', 'Communication' and 'Reciprocality' are reproduced in Figure 3.4.

Semantic frames can form hierarchies. More specific frames thus inherit properties from more general frames.

The FrameNet lexical database currently contains around 8,900 lexical units (a word in one of its senses) grouped into 625 semantic frames. Semantic frames are interlinked with their instances in 130,000 corpus sentences.

3.5 SALSA

The aim of SALSA (The Saarbrücken Lexical Semantics Annotation and Analysis Project)[2] is to provide a large, frame-based lexicon for German, with rich semantic and syntactic properties, as a resource for linguistic and computational linguistic research ([Erk et al., 2003]).

SALSA uses the FrameNet dictionary (the same set of frames, though developed for English) as the basis for its annotation. The goal of the annotation is to interlink the frames with their instances in the (syntactically annotated) TIGER corpus ([Brants et al., 2002]). So far, more than 20,000 instances have been finished. Annotated instances of 'Request' and 'Conversation' frames are depicted in Figure 3.5.

[2] http://www.coli.uni-saarland.de/projects/salsa/

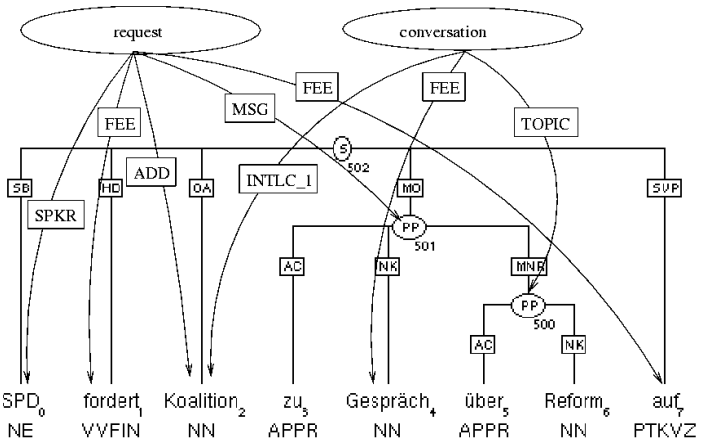


Figure 3.5: SALSA: annotated instances of FrameNet frames in the TIGER constituency tree.

See also [Ellsworth et al., 2004] for comparison of PropBank, SALSA, and FrameNet.

3.6 English Verb Classes and Alternations

The key issue of the description of verb behavior presented in [Levin, 1993] is the notion of alternation. Different alternations correspond to different changes of argument structure of lexical units. For instance, the verb *radiate* undergoes the substance/source alternation *Heat radiates from the sun – The sun radiates heat*), whereas the spray/load alternation couples can be illustrated on the pair *Jack sprayed paint on the wall – Jack sprayed the wall with paint* etc. Beth Levin recognizes around 80 alternations in English (the repertory of alternations is language dependent).

The main assumption of the work is that the behavior of a verb, particularly with respect to the expression and interpretation of its arguments, is influenced to a large extent by the meaning of the verb. Thus the set of alternations which are (or are not) applicable on the given verb should determine its semantic class. Or in other words, various combinations of applicable alternations should delimit semantically coherent verb classes.

This hypothesis was tested on around 3100 English verbs and led to 193 verb classes. The classes are described one-by-one in the work and the (non)applicability of individual alternations is substantiated for each of them. For instance, the "Build Verbs" (arrange, assemble, bake, blow, build, carve, cast, chisel ...) participate in material/product alternation (*Martha carved a toy out of the piece of wood/the piece of wood into a toy*), raw material subject alternation (*Martha carved beautiful toys out of this wood / This wood carves beautiful toys*), but does not participate in causative

alternations (*Martha carved a toy out of the piece of wood / *A toy carved out of the piece of wood*), etc.

3.7 PropBank

The main goal of the Proposition Bank project[3] (PropBank, see for instance [Kingsbury et al., 2002] or [Kingsbury and Palmer, 2002]) is to add a level of semantic annotation into the phrase-structure Penn Treebank[4] trees.

The PropBank annotation started with verbs. For each verb it was necessary to distinguish its individual senses first, and then to store argument structures of the verb in each sense separately. Arguments are distinguished only by numbers: Arg0, Arg1, ..., A5 (note that these numberings may have different meaning with different verbs: Arg2 in the frame of one verb can correspond to a completely different role than Arg2 in the frame of another verb). Modifiers are denoted as ArgM and further subclassified: ArgM-LOC (location), TMP (time), MNR (manner), DIR (direction), CAU (cause), NEG (negation marker), MOD (modal verb), EXT (extent), PRP (purpose), ADV (general-purpose modifier).

Penn Treebank data are annotated as follows: for each verb one of its senses is selected and its arguments and modifiers (in fact nodes in the phrase-structure tree) are marked with the above symbols (Arg0 etc.). The predicate is marked with 'Rel'. Simplified example of annotation:

```
He was drawing diagrams and sketches for his patron.
Arg0:      he
Rel:       drawing
Arg1:      diagrams and sketches
Arg2-for:  his patron
```

Besides such annotated treebank data, also so called Frame Files are created. They contain all senses of the processed verbs, and each sense is accompanied with a usage example and a list of arguments. Entry for the first sense of the verb *to go* is depicted in Figure 3.6.

3.8 Japanese-English valency dictionary

A large valency Japanese-English dictionary was built because of the machine-translation project ALT-J/E (Automatic Language Translator - Japanese to English) (see [Bond and Shirai, 1997]). The dictionary contains a semi-automatically created valency frames for 16,000 Japanese verbs, and a few more thousand have been added fully automatically. Each dictionary entry

[3] www.cis.upenn.edu/~ace/

[4] www.cis.upenn.edu/~treebank/

```
Roleset go.01 "motion":
 Roles:
  Arg1:entity in motion/goer
  Arg2:extent
  Arg3:start point
  Arg4:end point
  ArgM-LOC:medium
  ArgM-DIR:direction (usually up or down)

Examples:
 start and end points (-)
   What flights go from Seattle to Boston via Minneapolis?
     Arg1: what flights
     REL:  go
     Arg3-from:  Seattle
     Arg4-to:  Boston
     ArgM-LOC:  via Minneapolis

 extent (-)
   Imports have gone down 33%
     Arg1:  Imports
     ArgM-DIR:  down
     Arg2-EXT:  33%

 extent and end point (-)
   Woolworth went up 1 3/4 to 59 1/2.
     Arg1: Woolworth
     REL:  went
     ArgM-DIR:  up
     Arg2-EXT:  1 3/4
     Arg4-to:  59 1/2

 with direction (-)
   A lot of people would like TRACE to go back to 1970.
     Arg1:  TRACE -> a lot of people
     REL:  go
     ArgM-DIR:  back
     Arg4-to:  1970
```

Figure 3.6: PropBank entry for the first sense of the verb “to go”.

Japanese side	role	English side
iku$_1$		**go**$_1$
S1 ga agent, vehicle, animal	N1	NP Subj
S2 ni e made -road, -rail, theatre places, place	N3	PP *to* Acc
S3 kara yori -road, -rail, places, place	N4	PP *from* Acc
iku$_2$		**go**$_2$
S1 ga agent, vehicle, animal	N1	NP Subj
S2 ni e made -road, -rail, theatre places, place	N8	PP *along/around* Acc

Figure 3.7: Dictionary entry of the Japanese verb *iku*

contains a predicate, one or more frame elements (case slots) and an information about modality. On the Japanese side, each slot is accompanied with an information about its syntactic form and semantic constraints.

Syntactic form is defined as a phrase type: clause, noun phrase, or adverb. Particles[5] which can occur with the noun phrases are also stored in the noun slots.

Entries in both languages are interlinked via case roles. The following roles are used in the lexicon:[6] N1 (Agent), N2 (Object-1), N3 (Object-2), N4 (source), N5 (Goal), N6 (Purpose), N7 (Result), N8 (Locative), N9 (Reciprocal), N10 (Quotative), N11 (Material), N12 (Cause), N13 (Instrument), N14 (Means), QUANT (Quantity), TIME (Time), ADV (Adverb), TN1 (Time-position), TN2 (Time-source), TN3 (Time-goal).

Sample of the lexicon entry is depicted in Figure 3.7.

3.9 Smolensk database of verb features

The database studied in [Silnickij, 1999] results from a complex research running at the University of Smolensk since 1975. It contains verbs of ten languages: English, French, German, Russian, Armenian, Turkish, Arabic, Chinese, Indonesian and Japanese. There are 800-2500 verbs for each language in the database. Various features are stored for each verb (in each

[5]Particles are a closed class of postpositional case markers that mark Japanese noun phrases

[6]There are 24 different roles used in the ALT-J/E project, but only 14 of them are used in the lexicon.

language). Altogether, there are 64 features in the database, classified into the following groups:

- phonetic features
- morphological features
- etymological features
- chronometric features (in which period the verb first appeared)
- syntactic features
  - transitivity
  - the ability to bind an object in a morphological case different from accusative
  - ability to bind an obligatory adverbial
  - ability to bind a subordinating object clause
- semantic features
  - thematic features – they divide the verbs in three semantic macroclasses: energetic (related to transfer of material energy by motion, physical processes etc.), informatic (related to human processing of information), and ontological (for verbs requiring a higher degree of abstraction)
  - chronostructural features – three chronostructural verb classes are distinguished: processives, causatives, operatives.

The authors used the correlations in the set of features for building typologies of verb systems. Unfortunately no samples from the database are given in [Silnickij, 1999].

3.10 Valency dictionary of Bulgarian verbs

The valency dictionary [Popova, 1987] contains around 1000 most frequent Bulgarian verbs. For each verb, the following information is provided:

- transitivity and morphological aspect
- verbal description of the meaning of the verb
- I. frame - list of arguments required by the verb
- II. morphological information about the arguments
- III. semantic information about the arguments
- IV. examples of usage of the verbs

The authors describe the valency frame as an expression the first element of which is subject, then the predicate comes, then direct and indirect object, subordinating clause or adjunct. The morphological level specifies

ПРИНАДЛЕЖÀ, -ѝш, -ѝ, -àт; *несв.; непрех.*

1. Притежание съм в пряк или преносен смисъл на някого или нещо.

I. П+принадлежа+$О_2$.

II. П=С, М.

$О_2$=на+$С_1$, на+Мп, Мкд.

III. С=лице, предмет, идея.

$С_1$=лице, идея, дейност.

IV. Принадлежеше вече на Мария (Д. Димов). Принадлежат не на мъжа (П). То вече принадлежеше на всички (П). Принадлежеше на тебе (Ц. Цанев). Животът ѝ принадлежи на онова голямо нещо (С). . . . което по право ми принадлежи (Б. Райнов). Аз ти принадлежах дори без честна дума (П).

2. Спадам, числя се, отнасям се към нещо.

I. П+принадлежа+$О_2$.

II. П=С, М.

$О_2$=към+$С_1$, към+Мп.

III. С=лице, предмет, явление и др.

$С_1$=група, категория, система.

IV. Семейството принадлежеше към висшето дрезденско общество (А. Константинов). От тази точка най-добре можеше да се види към коя система принадлежи Мусала (Ив. Вазов, РСБКЕ). Той не принадлежеше към това племе (Е. Константинов). . . . към който принадлежи (Марксистко-ленинска философия).

Figure 3.8: Sample from the Bulgarian valency dictionary.

for instance that the given subject can be a noun or a pronoun, whereas the semantic level description may say that it is a person, animal, etc.

Sample from the dictionary is depicted in Figure 3.8.

The dictionary was recently converted into an electronic format, namely into XML ([Balabanova and Ivanova, 2002]).

3.11 Explanatory Combinatorial Dictionary of Modern Russian

Explanatory Combinatorial Dictionary (ECD) is a dictionary based on the Meaning-Text Model Theory (see [Mel'čuk, 1988] for references). A fragment of the ECD was published as [Mel'čuk and Zholkovsky, 1984] and contains dictionary entries for 250 Russian verbs and nouns. A regular ECD entry is divided into ten zones:

1. Morphological information (declension or conjugation type, aspect of verbs etc.),

2. Stylistic specification (archaic, colloquial, substandard etc.),

3. Definition, consisting of constants (elementary and derived concepts) and variables,
4. Government Pattern (see below),
5. Restrictions on the government pattern (conditions under which the actants of the entry lexeme can co-occur),
6. Examples - possible ind impossible (starred) combinations of the lexeme with its actants,
7. Lexical Functions - relations to other lexemes,
8. Illustration - the use of the lexeme and the corresponding LF in actual sentences,
9. Encyclopedic information (in a limited extent),
10. Idioms - list of semantically unanalysable idiomatic expressions in which the given entry lexeme appears.

Government pattern is a table in which each column represents one semantic actant of the lexeme (marked by the corresponding variable), and each element in the column represents one of the possible surface realizations of the corresponding syntactic actant. A part of a dictionary entry (including government pattern) is depicted in Figure 3.9 (the whole entry is 14 pages long).

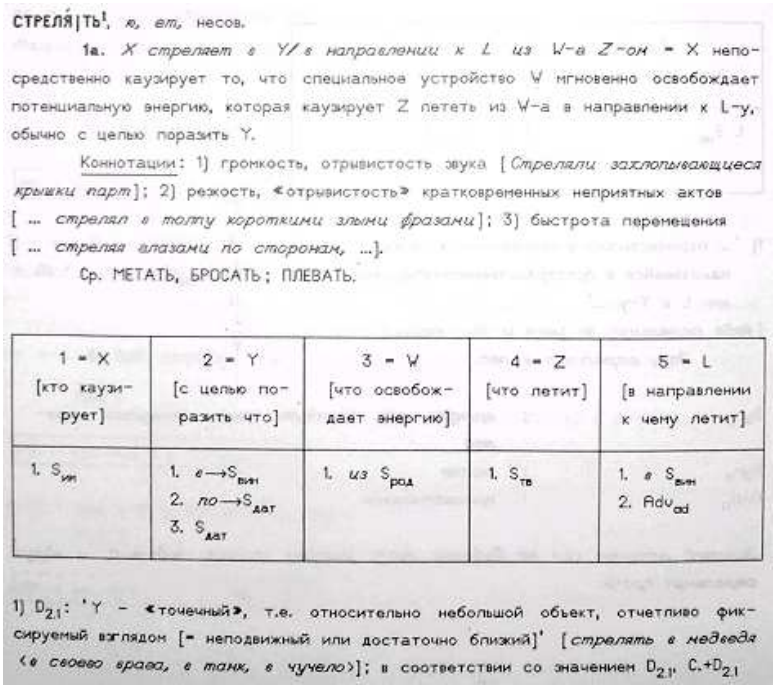
3.12 Dictionary in ETAP-3

ETAP-3[7] is an English-Russian machine-translation system developed at the Russian Academy of Sciences and based on the Meaning-Text Theory ([Mel'čuk, 1988]). It translates from Russian to English and vice verse. The translation dictionary used in ETAP-3 ([Boguslavsky et al., 2004]) is a successor of [Mel'čuk and Zholkovsky, 1984], although significantly changed. On one hand, the dictionary structure has been simplified so that the dictionary remains manageable even if it grew in the order of magnitude, but several new features have been added on the other hand in order to meet the needs of the MT system. The dictionary in ETAP-3 contains around 80,000 lexemes for each language.

A sample from the translation dictionary for English-to-Russian direction is depicted in Figure 3.10.[8] The lines starting with D describe the

[7] http://cl.iitp.ru/etap/index_e.html

[8] We would like to thank Leonid Iomdin for providing the sample from the ETAP-3 translational dictionary. The entry of the verb 'to distinguish' was by far the simplest and shortest one from that sample, since it contains neither complex translational rules nor lexical functions.



СТРЕЛЯ́|ТЬ[1], *ю, ет,* несов.

1а. *X стреляет в Y// в направлении к L из W-а Z-ом* = X непосредственно каузирует то, что специальное устройство W мгновенно освобождает потенциальную энергию, которая каузирует Z лететь из W-а в направлении к L-у, обычно с целью поразить Y.

Коннотации: 1) громкость, отрывистость звука [*Стреляли захлопывающиеся крышки парт*]; 2) резкость, «отрывистость» кратковременных неприятных актов [*... стрелял в толпу короткими злыми фразами*]; 3) быстрота перемещения [*... стреляя глазами по сторонам, ...*].

Ср. МЕТАТЬ, БРОСАТЬ; ПЛЕВАТЬ.

1 = X [кто каузирует]	2 = Y [с целью поразить что]	3 = W [что освобождает энергию]	4 = Z [что летит]	5 = L [в направлении к чему летит]
1. $S_{им}$	1. *в* $\rightarrow S_{вин}$ 2. *по* $\rightarrow S_{дат}$ 3. $S_{дат}$	1. *из* $S_{род}$	1. $S_{тв}$	1. *в* $S_{вин}$ 2. Adv_{ad}

1) $D_{2.1}$: 'Y – «точечный», т.е. относительно небольшой объект, отчетливо фиксируемый взглядом [= неподвижный или достаточно близкий]' [*стрелять в медведя (в своего врага, в танк, в чучело)*]; в соответствии со значением $D_{2.1}$, $C_i+D_{2.1}$

Figure 3.9: A part of the dictionary entry for 'to shoot' in [Mel'čuk and Zholkovsky, 1984].

surface forms of the verb arguments (only noun and prepositional groups are used in the given example). As it can be derived from the Russian zone of the entry, *to distinguish* is translated as *rozličat'* by default, but if the preposition *into* is used in the original English sentence, then the verb is translated as *podrazdeljat'* with preposition *na*, whereas *otličat'* and preposition *ot* is used if the preposition *from* occurred in the original sentence.

```
02495 23:03:50 05-05-2004          DISTINGUISH
      POR:V    АЛ
      SYNT:TRANSIT,SYM3-2
      DES:'ДЕЙСТВИЕ','ФАКТ','АБСТРАКТ'
      D2.1:S
      D2.2:BETWEEN1
      D3.1:FROM,NEMP
      D3.2:INTO,NEMP
      ******************************
ZONE:RU
      TRANS:РАЗЛИЧАТЬ
TRAF:TRADUCT2.29
LA:INTO,LR1:ПОДРАЗДЕЛЯТЬ,LR2:НА1
LA:FROM,LR1:ОТЛИЧАТЬ,LR2:ОТ
TRAF:RA-EXPANS.13
LA:FROM
```

Figure 3.10: Entry from the ETAP-3 translation dictionary.

3.13 The Proton valency dictionaries

There are two Proton valency dictionaries,[9] one for French (8500 entries, representing 3700 verbs) and one for Dutch (6299 entries for 4200 verbs) ([van den Eynde and Mertens, 2003]). Both databases provide an inventory of constructions in which a given verb can occur. The syntactic information for each verbal valency scheme is represented as the set of valency positions, each of which contains a list (paradigm) of (mainly) pronouns representing the possible instantiations of that valency position. Using the pronouns as representants of possible forms of valency positions is called pronominal approach ([den Eynde and Blanche-Benveniste, 1978]) and used by many other researchers.

A sample from the Proton dictionaries is presented in Figure 3.11.

3.14 Conclusion

We hope that a lesson can be learned from the wide survey of approaches given in the last two chapters. Since any attempt at a critical in-depth comparison would definitely go behind the limits laid on this thesis, we will at least summarize the main observations:

- There are huge differences in complexity of lexicon entries, probably with [Mel'čuk and Zholkovsky, 1984] being the extreme.
- Two methodologies can be distinguished in building the lexicon: *verb-wise* (the verb entries are completed one after another; applied in most printed dictionaries), and *frame-wise* (where verbs with given senses belonging to a certain frame are processed together at a time, e.g. in FrameNet). Both directions have their *pros* and *cons*: on the one hand, the danger of incompleteness has to be faced when using the frame-wise approach (some of the senses of the processed verbs never get described) as it was discussed in e.g. [Pustejovsky et al., 2004], whereas on the other hand the lexical resources created using the verb-wise approach tend to be more redundant because of disregarding some regularities. As [Briscoe, 2001] puts it: "Most grammatical frameworks treat valency almost entirely as a lexical property of predicates, although the inventory of valency frames ... can be described somewhat independently of individual words".
- In many recently developed resources, one can see the trend of interlinking (or even co-development) of lexicons with syntactically annotated corpora (e.g. PropBank, SALSA, PDT-VALLEX).

[Stevenson, 2003] says that "most NLP researchers do not want to spend time constructing their own lexicons since it is a difficult and time con-

[9] http://bach.arts.kuleuven.ac.be/PA/

VAL$ attraper
VERB$ attraper
PRED$ simple_predicator
ALPHA$ ATTRAPER
CLASS$ verb
NUM$ 9040
EX$ r : la police a fini par attraper le voleur
TR$ pakken, grijpen, vangen, beetnemen, betrappen, oplopen, (bus) halen, treffen
P0$ je, nous, on, qui, elle, il, ils, celui-ci, ceux-ci
P1$ te, vous, qui, ceci, la, le, les, en Q, que, celui-ci, ceux-ci, ça, l'un l'autre, se réc.
RP$ passif être, se passif, se faire passif
NEWEX$ Jean a pu attraper Paul

VAL$ attraper
VERB$ attraper
PRED$ simple_predicator
ALPHA$ ATTRAPER
CLASS$ verb
NUM$ 9050
EX$ r : que je t'y attrape
TR$ betrappen
P0$ je, nous, on, qui, elle, il, ils, celui-ci, ceux-ci
P1$ te, vous, qui, la, le, les, en Q, celui-ci, ceux-ci, l'un l'autre, se réc.
P2$?y, y(à_inf), ?quoi, ?ça, ça(à_inf)
RP$ passif être
NEWEX$ je l'ai attrapé à saboter la voiture de leur père
PIVOT$ p1, [in main clause is] p0, [in subclause for] à_inf, [in] p2

Figure 3.11: Sample entry from the Proton dictionaries: French verb *attraper* (to catch up).

suming process." However, in our opinion the core of the problem lies elsewhere: the growing diversity of lexical resources indicates that those of the researches who decide to build their own lexicon mostly do not spend much time by studying what was done elsewhere and thus the research is often isolated from achievements of other schools.[10] It can be also viewed as a consequence of what is (in a quite radical way) expressed in [Bolshakov and Gelbukh, 2000]:

> It is pleasing and profitable (and sometimes quite necessary for

[10]However, a few more detailed comparisons of at least two or three different approaches already appeared in the literature (e.g. [Hajičová and Kučerová, 2002], [Rambow et al., 2003] or [Ellsworth et al., 2004]), but they focus only on the most prominent projects in most cases.

getting a position in the university or for the project financing) to be author of a new fashionable theory. In the same time, it is necessary to be somewhat an altruist in our mercenary world, to consciously support someone else's theory and by everyday work to assist its promotion to the scientific circulation. It is easier to invent new title, terminology, and formalism without significant deviation from the "mainstream", to save comprehension on the side of those who had invented similar theoretical means in the past.

Chapter 4

Organizing the Lexical Space

Lisa: What a whimsical building.
Who says science can't be fun?
Bart: Me. I smell a museum.
Homer: Yeah, good things don't end with eum.
They end with mania. Or teria!
[The Simpsons]

Generally, the aim of any valency lexicon is to provide its user with the information about the combinatorial potential of . . . of what? Of words? The common-sense sense of the word "word" is notoriously vague to be used directly in a formal system. There is an abundance of more specific terms at our disposal, such as word form, lemma (with rarely used hyperlemma and superlemma), lexis, lexical unit, lexical item, lexical token, literal, lexeme (and occasionally seen hyperlexeme), lexia (completely unrelated to the medical term hyperlexia), headword, vocable, phraseme, multi-word expression, collocation. . . But again, unfortunately they are not used consistently in the literature. Therefore we have to explicitly state which terms (in which sense) have been chosen for describing our valency lexicon VALLEX. This is the first goal of this chapter.

Even though the form of a language sign is primarily arbitrary (in the Saussurean sense), properties of some lexical entities are often strongly motivated by properties of other lexical entities (Kuryłowicz's syntactic derivations be just one representant of such relations). Ignoring such relations would result in an increased redundancy of the lexicon. Thus the second goal of this chapter is to provide the valency lexicon with mechanisms that effectively reflect such regularities.

4.1 Lexemes, Lexical Units and Lemmas

Although lexeme is a fundamental concept of language description and lexicography in particular, it is often used without sufficient clarification. There are many informal definitions such as "lexeme is a minimal lexical unit of a language", which of course do not help much. One can find also more elaborated definitions of the notion of lexeme, e.g. in [Filipec, 1994]:

> The lexeme (. . .) is a formal-semantic unit of the lexical level in its intersection with the semantic level. Its status is deep and hierarchical. We distinguish the lexeme-type on the level of abstraction and the lexeme-token (allolex) on the empirical level in context use. This use is either untypical and individual or usual, typical and reproduceable. On the level of abstraction, there are three lexemic modes: the polysemic lexeme (hyperlexeme, for examples, "to give" with a set of meanings), the monosemic lexeme (hereafter lexeme: "to give something to somebody"), and the lexeme in typical context ("to give him a book, money, water to drink") . . . lexemes are realized by morphemes and phonemes and themselves realize sentences and texts.

This three-level definition is not broadly followed, however, it clearly illustrates the fact that lexeme is a mental construct, inevitably requiring certain abstraction. Although it is obvious that some abstraction is necessary and useful, one should always keep in mind that the way how language expressions (in given contexts) are grouped into lexemes is always only a matter of decision. The existence of lexemes as abstract entities cannot be proved or disproved and, moreover, there is no guaranteed true clue for determining lexeme identity (be the notion of lexeme leveled or not). Even if the tradition gives a clear answer in some situations (e.g. word forms 'gave' and 'gives' are always treated as instances of the same lexeme), it is less convincing in others ('given' and 'to give up'?).

However, even if there is no truth, it is impossible to build a lexicon without making any abstraction. Some 'clusterings' still seem to be better than others, e.g. due to some special linguistic, implementational, economic or aesthetic reasons.

Besides lexeme, we decided to use the notion of *lexical unit* (LU) in the following sense (citation from [Verspoor, 1997], page 216):

> Cruse (1986) distinguishes lexemes from lexical units. . . . The latter are form-meaning complexes with (relatively) stable and discrete semantic properties, and the meaning component is called a sense, corresponding to the intuitive notion of sense. . .

Obviously, the term LU is used here roughly in the sense of Filipec's 'monosemic lexeme'—loosely speaking, given word in the given sense. However, it should be distinguished that the term *lexical unit* cannot be interchanged with the term *sense*: the latter is only one of more components of the former.

Now back to the term *lexeme*. In our approach, lexeme is an abstract data structure that only associates lexical form(s) with lexical unit(s), nothing more. Or, in other words, lexeme is an ordered pair which couples the

set of lexical forms and the set of lexical units, as it is presented in Figure 4.1. Now, the term of lexeme becomes clear, all the possible questions were shifted to the terms lexical forms and lexical units.

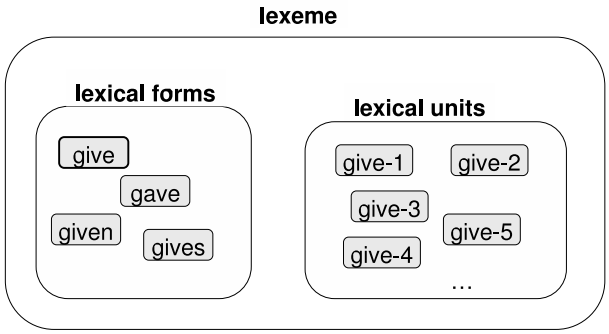


Figure 4.1: Illustration of the notions of *lexeme*, *lexical form*, and *lexical unit*.

In the language use, a lexeme can be manifested by lexical forms. The set of lexical forms of a lexeme contains virtually all manifestations of the lexeme (be they sequences of graphemes or phonemes). These manifestations are also called word forms. Note that the lexical form of a lexeme can be composed of several (even non-contiguous) parts of uterrance, such as in case of complex verb forms.

The set of lexical forms of a single lexeme can be quite large, especially in inflectionally rich languages. In most dictionaries, usually only one representant is chosen from the set (and possibly also its irregular forms, or also its inflectional paradigm), instead of listing all inflected forms. In case of verbs, infinitive is traditionally used in most Indo-European languages.[1] The selected representant is usually called *headword* or *lemma*. However, as it will be shown in the following sections, associating one lexeme with one such representant is not sufficient in our model (especially because of the fact that aspectual counterparts are to be considered as manifestations of the same lexeme and they can hardly be represented by one lemma).

Note that even the term 'lemma' is ambiguous in linguistic literature. Besides denoting the conventionally selected representant, it is also used for denoting the whole set of forms, such as at:

- http://www.searchmorph.com: "A lemma is a set of related morphological forms. These are related by orthography in most cases."
- or at http://www.concordancesoftware.co.uk/: "Lemmatising means grouping related words together under a single headword. For example, you

[1] There are exceptions such as Bulgarian, which does not have infinitives in its inflectional system and lexicographers traditionally use third person singular instead. Even in Czech there are word forms which are usually treated as verbs but which have no infinitive (*lze* – 'it is possible to').

> could choose to gather the words *am*, *was*, *are*, *is*, *were*, and *been* together under the word *be*. To use linguistic terminology, the variants taken together form the lemma of the lexeme *be*."

We do not use term 'lemma' in such set sense in this thesis.

Following the Czech tradition, we treat the expression *smát se* (lit. 'to-laugh refl-particle') as a single verb lemma. However, in some cases it is necessary to speak only about the verb part of the lemma (*smát* in this case), separated from the reflexive particle. We will use the term *m-lemma* for this purpose (*m* denotes the relation to the morphological layer of the Prague Dependency Treebank, where *smát* and *se* are two isolated tokens, see Section 5.2).

Conclusion from this section for the structure of VALLEX is the following: VALLEX will be composed of lexemes, which are abstract data structures associating one or more lemmas (and implicitely the whole set of their inflected forms) with one or more lexical units, corresponding to individual senses of the lemma(s).

4.2 Reflexive lexemes

In Czech, reflexive morphemes[2] *se*/*si* are used to express several different functions: either the morpheme is an obligatory formal component of the lexical form of a lexeme, or its presence is implied by grammar ((true) reflexivity, reciprocity, reflexive passivization ..., see [Skoumalová, 2001], pages 21-33 for detail).

If the reflexive morpheme is obligatorily present in all lexical forms of the lexeme, then it is traditionally treated as a reflexive particle[3] (whereas in case of true reflexives it has the status of a reflexive pronoun). This happens in two cases:

- *reflexivum tantum*—there is no irreflexive counterpart of the verb, typical examples are *smát se* (to laugh) and *bát se* (to fear/to be afraid). There is no reflexivity in the meaning of these lexemes.
- *derived reflexive*—these lexemes are derived from the irreflexive origins, but the lexical meaning is so distinct that it cannot be treated as a reflexive form of the original lexeme and rather constitutes a separate verb lexeme. Examples: *hodit* (to throw) / *hodit se* (to suit), *chovat* (to breed) / *chovat se* (to behave).

[2]The reflexive morphemes are always manifested by separated graphemes in Czech orthography, unlike in Russian (*nazyvat'sja*/to be called) or Spanish (*llamarse*/to be called).

[3]There are two exceptions possibly violating the obligatory presence of the particle: question context (*Smál se? Smál.*), and haplology (haplology is the loss of one of two identical or similar adjacent syllables; *Chce se mi smát* instead of *Chce se mi smát se*).

In case of reflexiva tantum, no decision-making related to separating the reflexive lexeme is necessary, as no irreflexive lexeme is possible. However, in case of derived reflexives, the criterion of sufficiently different lexical meaning is of course not reliable. Although broad attention was paid to reflexives in the linguistic literature, we have not found any clear-cut solution for distinguishing instances of derived reflexives from true reflexives.Thus it is often difficult to state whether a separate reflexive lexeme should be encoded in the lexicon or not. Besides the clear cases such as the above examples (which are unfortunately rather an exception than a rule), there are situations when the common meaning core can be easily traced back, e.g. *řídit* (to direct) / *řídit se* (to adhere to), and it is not clear whether the semantic divergence is sufficient for establishing a separate lexeme.

The fuzzy border between the reflexives and irreflexives is probably due to the fact that lexicalization of reflexives is a continuous process in the language (even in case of reflexiva tantum: [Tabakowska, 2003]: "On diachornic scrutiny, most reflexiva tantum prove to be, in earlier stages of language, semantically (and formally) transitive.").

Conclusion for VALLEX is the following:

- there might be pairs or tripples of different lexemes sharing the same m-lemma (e.g. *brát / brát se / brát si*),
- LUs of irreflexive lexemes should be ideally equiped with information specifying which types of reflexivity are applicable on the given LU.

It is an open question what is the optimal lexicographic solution in the rare cases where the presence of a reflexive particle seems to be optional (i.e., it can be omitted without changing the meaning, at least in some contexts). Examples: *kamarádit s někým* vs. *kamarádit se s někým* (to be friend with someone), *myslet, že...* vs. *myslet si, že...* (to think that...), *končit* vs. *končit se* (to end). In the presented version of VALLEX, they will be represented as separated lexemes for the sake of consistency.

4.3 Lemma variants

Lemma variants (also spelling variants, orthographic variants, or doublets in [SSJČ, 1978]) are groups of two or more lemmas that are interchangable[4] in any context without any change of the meaning, e.g. *dovědět se / dozvědět se* (to learn) in Czech or *color / colour* in English. The only difference usually is just a small alternation in the morphological stem, which might be accompanied by a subtle stylistic shift, such as in the case of *myslet / myslit* (to think), where the latter variant is rather bookish. Moreover, although the infinitive forms of the variants differ in spelling, some of their

[4]Though the definitions seem to be similar, the term 'lemma variants' should not be confused with the term 'synonymy'.

conjugated forms are often identical (*mysli* (imper.sg.) for both *myslet* and *myslit*).

In Czech, the main sources of lemma variants are the following:

- changed orthography of borrowings from foreign languages, such as in *organisovat / organizovat* (to organize),
- stem variants such as in *chytit / chytnout* (to catch), or *rozpočíst / rozpočítat* (to divide),
- vocalized and non-vocalized variants of prefix morphemes for verbs created by prefixation, such as in *odjet / odejet* (to leave), or *podpisovat / podepisovat* (to subscribe).

Archaic forms of infinitives ending with *-ti* could be theoretically treated as lemma variants too (they bear the same lexical meaning although there is a difference in spelling), but there is no need to list them in a lexicon, since they can be formed for vast majority of verbs (and similarly with *-ci*, e.g. in *říci / říct* (to say), which seems to be applicable on all verbs ending with *-ct*). Note that this type of variation can be naturally multiplied with some other types (*podpisovat / podpisovati / podepisovat / podepisovati*).[5]

In some situations, it is not easy to decide whether two lemmas are variants or not, e.g. *ovázat / obvázat* (to bind, to bandage).

Conclusion for VALLEX is that in the case of lemma variants, both (or more) lemmas should be explicitely stored in the lexeme entry.

4.4 Homographs

Homographs are lemmas ("accidentally") identical in the spelling but considerably different in the meaning (there is no obvious semantic relation between them and origin).[6]

In lexicography, homographs are traditionally treated as different lexemes. In the following paragraphs, the homographs will be distinguished using Roman numbering in superscript.

It has been pointed out many times in the literature on lexical semantics that a clear operational distinction between homography and polysemy is lacking. Obvisously, there is no reliable measure of the difference in meaning that would distinguish homography from polysemy, and as for the word origin, there is the danger of amateur etymology ([Kilgarriff, 1992], page 46).

[5] However, some other combinations might signal naive stylistic attempts at a solemnity of the utterance, e.g. the combination of modern spelling of a borrowed word with the archaic ending (e.g. organizovati).

[6] In Czech linguistic literature, the term 'homonym' is mostly used to express the same notion. However, in this aspect we adhere to the English tradition, where the term 'homograph' prevails and the term 'homonym' is often used in different meaning: a word pronounced the same as another, but spelled differently (which is called 'homophone' by others).

In VALLEX, the following types of homographs are captured:

- homographs differing in verbal aspect, e.g. imperfective *stačit*^{I} (to be enough) vs. perfective *stačit*^{II} (synonymous to *stihnout*, to catch up with)
- homographs differing in the set of (conjugated) lexical forms, e.g. *žilo* (past.sg.fem) for *žít*^{I} (to live) vs. *žalo* (past.sg.fem) *žít*^{II} (to mow),[7]
- homographs differing in etymology, e.g. *nakupovat*^{I} (to buy) vs. *nakupovat*^{II} (to heap). The difference is etymology is often confirmed by different aspectual counterpart, e.g. *opírat*^{I} / *oprat* (to wash) vs. *opírat*^{I} / *opřít* (to support),

Note that homography should be ascribed rather to m-lemmas than to lemmas: *dít*^{I} (to say, archaic) and *dít*^{II} *se* (to happen) have different etymology and cojungation patterns.

More than two lexemes can be related by homography. The example tripple is *stát*^{I} (to stand, to cost) / *stát*^{II} *se* (to happen, to become) / *stát*^{II} (to melt down, very rare). The noun *stát* (state) would be traditionally treated as their homograph too, but as the part of speech should be specified for each lexeme in VALLEX anyway, we find it redundant to distinguish it by an extra homograph index.

There are more sources of homography in Czech (e.g. verbs composed of different stems and different prefixes, accidentally resulted in identical lemmas, e.g. *od-rolovat* (to roll away) and *o-drolovat* (to crumble gradually), but they are too rare to occur in VALLEX. However, two "verbs" which are derived from the same lexeme by application of a polysemous prefix are not considered to be homographs in VALLEX: e.g. *vy$_1$-jít* (to go out) and *vy$_2$-jít* (to climb up) are treated only as different LUs within the same lexeme.

Obviously, the sources of homography are different in different languages. For instance, homographs differing in part of speech are very frequent in English (e.g. heap/to heap), but relatively rare in Czech.[8] This difference is probably caused by the fact that Czech has much richer word-formative morphology. Sources of some types of homography might be completely absent in other languages: for instance, in English or Czech there is no analogy of the difference between separable and inseparable prefix in German *umschrebein* (to paraphrase) / *um—schreiben* (to rewrite).

Conclusion from this section for VALLEX is two-fold:

[7] Note that the types of homography that occur in the lexicon are influenced by the word form representant which we have chosen to be the lemma. If we would have chosen 3rd.sg.past as a lemma instead of infinitive, there would be no homograph for *žít* in the lexicon, whereas if we would have chosen 3rd.sg.fut, then *růst* (to grow) and *porůst* (to overgrow) would have to be treated as homographs, although now they are not.

[8] At least when regarding infinitives. Example: the verb *růst* (to grow) and the noun *růst* (growth).

- additional symbols must be used to distinguish homographic lexemes, since their lemmas are identical in spelling,
- the symbols should be ascribed directly to m-lemmas.

4.5 Aspect

Perfective and *imperfective* verb forms are distinguished between in Czech; this characteristics is understood as grammatical, although exhibiting certain features of lexical derivation, and is called aspect.[9]

Within imperfective verbs, there is a subclass of *iterative* verbs. Traditionally, Czech verbs are claimed to form aspectual pairs (or triads, when counting separately also the iteratives), where the two (or three) counterparts share almost the same lexical meaning and differ mostly only in terminativity of the denoted process (or in another feature related to aspect). Example: perfective *dát*, imperfective *dávat*, iterative imperfective *dávávat*.

Some verbs can be used in different contexts either as perfective or as imperfective, for instance *informovat* (to inform) or *charakterizovat* (to characterize). They are called *biaspectual*.

There are three ways how the aspectual pairs are formed in Czech (sorted according to productivity):

- *suffixation*: imperfective verb is derived from the perfective one, e.g. by infix *-ova-*: *vypsat* / *vypisovat* (to excerpt, to write off)
- *prefixation*: perfective verb is derived from the imperfective one by adding a prefix: *psát* / *napsat* (to write)
- suppletive (phonemically unrelated) couples: *vzít* / *brát* (to take).

In most cases, more prefixed verbs can be derived from one base verbs. The decision is to be made which of them is the aspectual counterpart. For such case, the secondary imperfectivisation test can be used, the underlying idea of which is that if you cannot undo the perfectivisation process in any other way than by going back to the original basic verb, then you have formed a genuine aspectual pair. For instance, for *napsat* derived from *psát* there is no other imperfective different from *psát*.

However, in some cases it is difficult to select the prefixed aspectual counterpart of a given verb. First, it might happen that there are two prefixed verbs, each of them being the counterpart of the base verb only in one of its senses, e.g. *vyblednout* (to fade out) and *zblednout* (to turn pale) for *blednout*. Second, the secondary imperfectivisation test is not hundred percent reliable: even in the case of *napsat*, one can find the imperfective *napisovat* in [SSJČ, 1978]. Third, in some situations it is not clear whether the secondary imperfection test can be applied or not, since the secondary

[9] A huge databaze of 8000 titles on grammatical aspect and related topics (*Aktionsart*, tense etc.) can be found at http://www.scar.utoronto.ca/~binnick/TENSE/.

imperfective is formally derived from the prefixed verb, but as for syntactic properties (or the set of senses), it is similar only to the base verb, not to the prefixed one. Example: *žádát* (to beg, to require) / *požádat* (to beg) / *požadovat* (to require).

In our opinion, the only possible clear-cut solution is not to represent the prefixed verbs as the aspectual counterparts of the base verbs at all in the lexicon. This approach is thoroughly applied in VALLEX.

In VALLEX we follow FGD in considering aspectual counterparts to be just different forms of the same lexeme, see [Panevová et al., 1971].[10] Again, correctness of such approach cannot be "linguistically" proved or disproved. It has its pragmatic *pros* and *cons*, but from the empirical point of view the advantages seem to prevail.

The main advantage is that treating an aspectual pair as a single lexeme significantly reduces redundancy of the lexicon, since the aspectual pairs generally tend to share most of their senses (and syntactic properties too).[11]

However, there are exceptions: for instance the verb *odpovědět* (to answer) is the aspectual counterpart of the verb *odpovídat* (to answer, to be responsible, to mach) only in one of its senses (LUs). In other examples, interchanging the aspectual counterparts in a given context might result in semantically completely unrelated utterances:

(11) Pořádně sebou hoď!
Look alive!

(12) Pořádně sebou házej!
Trash about!

So the conclusion for VALLEX is that aspectual counterparts are to be merged into the same lexeme (and thus the set of lexical forms has to be represented by (at least) two lemmas), and that VALLEX has to be equiped with a mechanism that allows to limit the lexical forms of a given LU only to conjugated forms of only one of the counterparts.

4.6 Verb Determination

The perfectiveness/imperfectiveness opposition in Czech is related to the older opposition of determinacy/non-determinacy – the creation of the former was probably conditioned by existence of the latter, see [Kopečný, 1962] page 12-15. Today, there is only a small number of verb couples where this distinction occurs (i.e., has not changed into the aspect distinction): *jít / chodit* (to go), *jet / jezdit* (to go/ride/drive), *vézt / vozit* (to cart), *hnát /*

[10] On the tectogrammatical level of FGD, the aspect is represented not by the lexical value, but by a special grammateme.

[11] This claim is supported also by the fact that when translating from Czech to English, the aspectual counterparts are usually translated using a single English verb (the difference in aspect is to be expressed by different language means).

honit (to chase), *běžet / běhat* (to run), *letět / létat* (to fly), *nést / nosit* (to carry), *vést / vodit* (to lead), *táhnout / tahat* (to pull). But although they are not numerous, they have high relative frequencies in texts.

Similarly as in the case of aspect, although the paired verbs are different in their outer shapes, they share most of the lexical meaning (and as it is shown above, the pairs can be translated using the same English verb). That is why these pairs seem to be the next candidates to be merged into shared lexemes, but at the present stage of VALLEX we decided not to do so. The main reason is that it would require further significant complexification of the lexicon structure, which could be however utilized only for a very small amount of verbs.

4.7 Prefixing

Let us start with the citation from [Uher, 1987]:

> Verbal prefix is statutorily described as a *polyfunctional polysemous morpheme of an agglutination type* with its own specifically derivational function of word formation (modification), with a *concurrent* binding perfectivization (aspectual) function, and with other, limited functions. It is the decisive means in verbal determination. ... From the formal point of view prefixes represent a relatively closed inventory.

There are twenty basic Czech verb prefixes (not counting the vocalized versions): *do-*, *na-*, *nad(e)-*, *o-*, *ob-*, *od-*, *po-*, *pod(e)-*, *pro-*, *pře-*, *před(e)-*, *při-*, *roz(e)-*, *s(e)-*, *u-*, *v(e)-*, *vy-*, *vz(e)-*, *z(e)-*, *za-*.

Some base verbs are compliant with almost all of them, especially the verbs of motion: example: *jít* (to go), *dojít* (to run short), *najít* (to find), *nadejít* (to outgo), *obejít* (to go around), *odejít* (to leave), *pojít* (to die), *podejít* (to go under), *projít* (to go through), *přejít* (to pass over), *předejít* (to forewent), *přijít* (to come), *rozejít (se)* (to separate), *sejít* (to go down), *ujít* (to bilk), *vejít* (to go in), *vyjít* (to go out), *vzejít* (to arise), *zajít* (to perish).

Obviously, it is not desirable to merge a base verb and all its derived prefixed verbs into one lexeme, as the prefixes change the lexical meaning (sometimes the meaning even seems to be completely unrelated to that of the base verb). But we find it important to interlink the base with its prefixed derivates in the lexicon, as one can thus study new types of regularities in the lexicon (new in the sense that they cannot be empirically studied without having such interlinking in the lexicon). We hope that such regularities can be once used for an automatic prediction of syntactic (and also semantic) properties of lexemes which have not been registered by annotators yet.[12]

[12]However, such automatic prediction would require a module that detects what is a given verb derived from, which is not a trivial task. For instance, the verb *svítat* (to dawn)

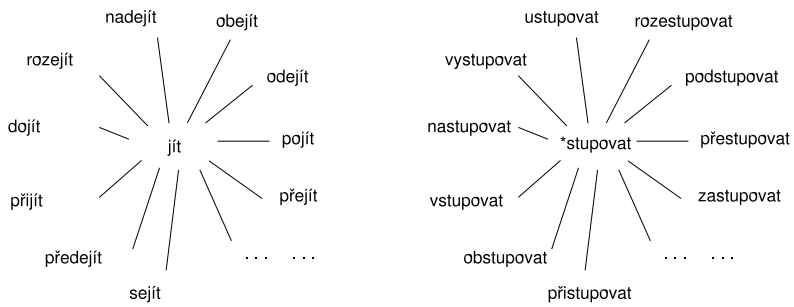


Figure 4.2: Examples of hub clusters formed by a (possibly non-existing) base verb and derived prefixed verbs.

For instance, verbs of motion derived by the prefix *při-*, such as *přijít* (to come), *přiletět* (to come flying), *přiběhnout* (to come running) etc., have the obligatory directional free modifier DIR3 (direction-to) in their valency frames. Such automatic prediction would be more successful in the case of highly productive prefixes, such as *na-* (plus reflexive particle *se*) in the sense of spending a lot of time by doing the activity denoted by the base verb.

It should be also noted that the relations rendering the verb prefixation should ideally interlink LUs in different lexemes and not the whole lexemes. For instance, *přijít* is derived from the primary LU of *jít* (to go) and has not to do with the other LUs (e.g. the sense 'to function' or 'to suit').

In some cases, a group of prefixed verbs looks as if derived from the same base verb, but the base verb does not exist in the language, as illustrated in the right part of Figure 4.2. Thus it is clear that it is not possible to represent such "hub" structure by listing all derived verbs in the entry of the base verb, nor by storing the base verb in the entries of prefixed verbs, because the base verb might simply not be present in the lexicon. Alternative data representation is following: the hub structure could be represented as a set of LUs (or rather pointers to LUs) with one (possibly absent) distinguished element representing the base LU.

Finally, it should be recalled that verb prefixing is not used only for deriving new lexemes in Czech, but also for expressing the future tense, such as in *běží* (lit. he/she runs) and *poběží* (lit. he/she will run). In rare cases, this might result in ambiguity: *poroste* can be either of form of *růst* (to grow) or *porůst* (to cover by growing).

At the present stage, the relations between prefixed verbs and their bases are not explicitly captured in VALLEX yet.

---

is etymologically unrelated to the base verb *vítat* (to welcome).

4.8 Negative lexemes

Negated forms of Czech verbs are formed using the *ne-* prefix, which also occurs in case of nouns, adjectives and adverbs. However, there are examples of nouns, adjectives and adverbs, which formally resemble negated forms (and also have certain negative component in their lexical meaning), but in a lexicon they must have their own lexeme because of one of the following reasons:

- the meaning of the prefixed word form is completely unrelated to the meaning of the unprefixed word form: *nesmyslný* (nonsensical) has not to do with *smyslný* (lascivious), or *nerudný* (grumpily) with *rudný* (ore);
- the positive unprefixed form does not exist in Czech at all: e.g. there is no positive *kalý* for *nekalý* (unfair).

See e.g. [Eisner, 2002] for more examples.

Surprisingly, similar examples are extremely rare among verbs. We are aware only of *nenávidět* (to hate) and *nesnášet* (to loathe). In the first case, the positive counterpart *návidět* is documented in 19th century literature ([SSJČ, 1978]), but it sounds fully unnatural to the contemporary speakers and the given lexeme is lemmatized as *nenávidět* in the contemporary lexicons. In the second case, lexeme *snášet* (to suffer, to lay, or the homograph in the sense to bring together) exists, but none of its meanings is the opposite to the meaning of the lexeme *nesnášet*.

There are also rare cases where only the negated verb forms can be used within certain idiomatic constructions, for instance in *Nedej se!* (Don't give up!). However, we do not find it necessary to introduce a new lexeme (with negated lemma) in VALLEX in such situations, as the semantic relation to the original unnegated meaning can be traced back. The fact that the verb must be negated in such usage will be captured by a different means in VALLEX (see Section 5.5).

In FGD, negation with verbs is understood as a specific lexical item in the tectogrammatical sentence representation.

The conclusion for VALLEX is that verbs (unlike nouns, adjectives, and adverbs) do not require a special treatment of negative lexemes.

4.9 Valency across parts of speech and lexical functions

VALLEX focuses only on verbs in the present version. However, from the long-term perspective, study of valency of other parts of speech is important too. Similarly as in the case of prefixation, it is important to study the relation between verbs and derived nouns, adjectives, or adverbs, because it can be potentially used for a (semi-)automatic expansion of the

lexicon. Again, the more productive is the word-formative process in question, the more regular will be the transformation of its valency frames. In [Lopatková et al., 2002] we have reported a preliminary experiment in which valency frames of Czech nouns ending with *-ní/-tí* derived from verbs were automatically predicted with the success rate around 70 %.

However, prediction of valency frames of derived lexemes can hardly be done fully automatically, for instance due to the fact that usually only some of the senses of the original verb are present with the derived lexeme. Thus, e.g. the noun *odpověď* (answer) is related to the verb *odpovídat* only in the sense 'to answer' or 'to react', whereas the noun *odpovědnost* is related to the same lexeme in the sense 'to be responsible'.[13] Therefore the derived lexemes, as well as the links between the basic and derived lexical units, should be ideally stored in the lexicon. We support this claim by two more examples:

- in most cases, valency frames of adjectives derived from verbs can be predicted with high accuracy; however, there are exceptions: for instance, the adjective *urozený* (noble) is both semantically and syntactically unrelated to the verb *urodit se* (to crop), although they are related from the word-formative point of view,

- it can be hardly predicted whether idiomatic constructions can be applied also on the derivatives: for instance the noun *příchod* (comming, arrival) related to the verb *přijít* (to come) can be used in no sense close to *přijít k sobě* (come to life).

In our opinion, the most promising way to capture the above mentioned relations in the lexicon is to use the framework of *lexical functions* (LFs) introduced within the Meaning-Text Theory (MTT, see for instance [Mel'čuk and Zholkovsky, 1984]). LFs provide a powerful formalism for representing relations between what we call lexical units (and what the authors of MTT call lexemes). LFs are functions in the mathematical sense: each lexical function returns an output value (lexical unit or a set of lexical units) for a given input value (lexical unit). There are tens of different LFs for various purposes defined in MTT, see e.g. [Wanner, 1996] for their survey. In the following list, we will try to illlustrate only a few LFs which return different types of syntactic derivatives for the verb *stavět* (to build):

[13] The need for a proper representation of relations between individual senses of derivationally related words was noted also in [Fellbaum and Miller, 2003] (they call such relations 'morphosemantic links'): "None of the traditional dictionaries we know provides a full mapping between all derivationally related word forms and senses. For example, the verb *write* is usually accorded numerous sense, whereas the noun *writer* may be dismissed with the gloss, "one who writes". But someone who writes novels is different from someone who writes music, and these writers in turn must be distinguished from someone who writes a letter or a computer program."

- $\mathrm{S}_{0}(stavět) = \{stavení,\ stavění\}$
- $\mathrm{S}_{1}(stavět) = \{stavitel,\ stavebník\}$
- $\mathrm{S}_{2}(stavět) = \{stavba,\ stavení\}$
- $\mathrm{S}_{loc}(stavět) = \{staveniště\}$
- $\mathrm{A}_{0}(stavět) = \{stavební\}$
- $\mathrm{A}_{1}(stavět) = \{stavící\}$
- $\mathrm{A}_{2}(stavět) = \{stavený,\ stavěný\}$

In the present version of VALLEX, the concept of lexical functions has not been implemented yet.

Chapter 5

Valency in Dependency Trees

To the great tree-loving fraternity we belong.
We love trees with universal and unfeigned love,
and all things that do grow under them or
around them—the whole leaf and root tribe.
Henry Ward Beecher

5.1 Valency and syntactic structures generally

Valency can be informally described as the combinatorial potential of language units. It is not a phenomenon directly observable in language – we can observe only its manifestations in concrete utterances (they might be grammatical or non-grammatical, incomplete, sounding odd in the given context ...), not valency itself! Any given verb can occur in infinitely many sentences, but we believe that the description of its valency can be discrete and finite. Therefore it is clear that the study of valency requires the ability of a huge generalization over the language performance.

Obviously it would not be wise to start with the generalization from the very bottom – e.g. by re-inventing declension patterns etc. We want to use as much from the theoretical background developed by the previous generations of linguists as possible. However, especially at higher "levels" of language description, we often have to choose one of many possibilities. The choice has to be made and the decision may have crucial consequences for our new framework.[1]

[1] As in any other science, some linguistic theories (hypotheses, models, data representations etc.) are better than others in that they are more exact, more adequate, more economic, more ellegant, simpler for understanding, with better explanatory power, with better predictions etc. In this aspect, we do not believe in "theory neutrality", which is a popular key-word in the present computational linguistics. Even such a basic thing as parts-of-speech classification should be viewed as a deliberetaly chosen solution (one of many possible), since it is a product of historical coincidence rather than a self-evident fact (it is well known that the traditional POS classification lumps together classes of words which have little or nothing in common and which simply cannot be sensibly forced together). Finally, we find it utmost surprising if some authors declare theory neutrality of their framework and then – without batting an eyelid – adhere to phrase-structure trees as to something given in the very same paragraph.

Human language, as an object of scientific interest, is extremely intricate. To decompose the task of its study into smaller parts, many linguistic theories postulate several levels (strata) of language representation, and transformation or other transducing procedures between the levels. Such accounts are often called stratificational, as opposed to monostratal accounts, in which each sentence has a single complex representation.[2]

Distinctions among the levels are a matter of continuing debate. In [Allen, 1995] the following levels are distinguished: (1) phonetic and phonological, (2) morphological, (3) syntactic, (4) semantic, (5) pragmatic, (6) discourse.

Representatives of multistratal dependency-oriented approaches are Functional Generative Description, the levels of which are enumerated in 2.1.2, and Meaning-Text Theory (MTT, see [Kahane, 2003] for a brief introduction and more references) with seven levels: (1) surface-phonological, (2) deep-phonological, (3) surface-morphological, (4) deep-morphological, (5) surface-syntactic, (6) deep-syntactic, (7) semantic.[3]

As it was said above, we can observe valency only via its manifestation in concrete sentences. Perceiving the sentence as a mere sequence of word forms would make the above mentioned generalization very complicated, since the form of the sentence is influenced by many phenomena having nothing to do with valency (in Czech, e.g., word-order rules for clitics, vocalization of preposition, and many others). Therefore it is extremely advantageous to work with syntactic structures instead.[4]

Nowadays, two main different types of syntactic structures are used for formal description of sentence syntax: (i) phrase-structure trees, based on immediate constituency, and (ii) dependency trees, based on asymetric head-dependent relations.

We use dependency trees in this thesis. Not only because of the long decades of Praguian tradition of dependency syntax (some references can be found in [Hajičová and Sgall, 2003]), but also because of the two following reasons:

- Dependency syntax allows for a more natural view on valency frames: valency frames can be seen as simple underspecified dependency trees, and thus the formal representation of a valency frame nicely matches the syntactic tree containing an instance of the frame; linking valency frames with phrase-structure trees necessarily results in drawing two

[2] But even in monostratal approaches, such as Head-Driven Phrase Structure Grammar, notions from different levels of language description are distinguished (see e.g. [Manning and Sag, 1998]), though in HPSG no level of grammatical knowledge is priviledged with respect to others, and no level is derived from any other.

[3] A short comparison of MTT and FGD can be found in [Žabokrtský, 2005].

[4] As Pustejovsky puts it, "without an appreciation of the syntactic structure of a language, the study of lexical semantics is bound to fail" ([Pustejovsky, 1995], p.5).

different types of trees over one another, an instant of which can be seen in Figure 3.5 on page 21.

- this thesis is primarily focused on Czech verbs, but phrase-structure formalisms perform poorly in Czech (as in any other languages that do not use constituency as the main surface expressive device; English is quite extreme in this aspect); moreover, the dependency approach has been succesfully verified for Czech on large data sets (thousands of syntactic trees in PDT).

Those who are interested in the "competition" between phrase-structure and dependency syntax can find the analysis of four (mostly extra-linguistic) reasons of the long-time dominance of phrase-structure approaches analyzed in [Mel'čuk, 1988], pages 3-7. However, certain convergence of the (Western) main-stream linguistics with the dependency-based approaches can be observed in the last decade. It can be illustrated on the recent developments in the field of treebanking. In [Kingsbury and Palmer, 2002], a layer of predicate-argument structures was added to English Penn Treebank – they started by marking clause nuclei composed of verbal predicates and their arguments. In fact, these structures can already be viewed as small (still rather flat and isolated) dependency trees, headed by (potentially complex) verbal forms. Later, they added also 'modifiers of event variables' (e.g. [Babko-Malaya et al., 2004]), thus broadening the nuclei of dependency trees with what FGD would call free modifiers. Then also the argument structures for instances of common nouns were added ([Meyers et al., 2004]). Finally, the isolated islands containing small dependency trees were connected with relations representing subordinate and coordinating conjunctions ([Miltsakaki et al., 2004]), thus forming a deeper structure, extremely similar to deep-syntactic (resp. tectogrammatic) dependency trees available in MTT and FGD decades ago.

However, many linguists are still not fully aware of the advantages offered by dependency structures. This can be shown on [Burchardt et al., 2005]: "Frame semantic annotations of contiguous texts are ... necessarily partial. Due to the missing constructional 'glue' in semantics composition, argument and variable binding cannot be defined in a strictly compositional way, and we obtain partially connected graphs of frame structures." – In dependency approaches, frames are nothing else than subgraphs of bigger dependency trees the natural property of which is the connectivity of the sentence representation. And one more citation from the same article: "A challenge in using frame semantic annotations as a partial text meaning representation structure is to produce more densely connected structures of frames by inducing co-reference relations between frames and frame roles." Also the coreference relations can be easily added to dependency trees, as it has been done for a large amount of data in PDT 2.0 (see e.g. [Kučová and Žabokrtský, 2005]).

5.2 Surface and deep syntactic trees in PDT 2.0 style

In the following sections we work with dependency trees in the style of Prague Dependency Treebank 2.0, based on the theorectical framework of Functional Generative Description. That is why we sketch (at least very briefly) the basic principles of its layered annotation scenario here (for more information, see the PDT 2.0 documentation).

In PDT 2.0, there are four layers of sentence representation: w-layer, m-layer, a-layer, and t-layer.

W-layer (word layer) contains the original sentence represented as a sequence of tokens (words or punctuation marks). No linguistically relevant information is added on this layer (even the errors present in the original text are preserved here).

On m-layer (morphological layer), the sentence is represented as a sequence of tokens too, but morphological tags and lemmas are added to each token (and the errors are corrected).

A-layer (analytical layer) captures a view of the surface syntax of the sentence. The sentence is represented as a rooted tree, in which each node corresponds to one token of m-layer (the only exception is the technical root of the sentence, having no m-layer counterpart). Each node is equiped with its analytical function,[5] which renders the surface-syntax role of the node within the sentence (the possible values are Pred - predicate, Sb - subject, Obj - object, Atr - attribute, Adv - adverbial etc.).

On t-layer (tectogrammatical layer), every sentence is represented as a rooted tree with labeled nodes and edges too, but the labeling is much more complex in comparison to a-layer. Only autosemantic words (nodes bearing a lexical meaning) have nodes of their own on t-layer, whereas functional words (such as prepositions, conjunctions or auxiliary verbs) have not (as it is shown in Figure 5.1, complex verb forms or prepositional groups "collapse" to single nodes on t-layer). Each t-layer node has a functor (tectogrammatical function) capturing deep-syntax role of the node with respect to its governing node (such as ACTor, ADDRessee, PATient, various types of temporal and location modifiers etc.) and t-lemma (tectogrammatical lemma) corresponding to the autosemantic lexeme in question, or containing a ficitious lexeme (#PersPron for personal pronouns, #Cor in case of unexpressed actor of infinitive verbs, #EmpNoun for an unspecified elided noun, and many others). Also attributes capturing topic-focus articulation, co-reference, morphological meanings (so called grammatemes, such as number for semantic nouns, degree for semantic adjectives, tense for semantic verbs) etc. are attached to the nodes where appropriate (according to node classification).

[5] However, the analytical functions should be understood as attached to edges rather than to nodes, as they represent the relations between node pairs.

As it is illustrated (in a simplified way) in Figure 5.1, all four layers are interlinked by pointers.

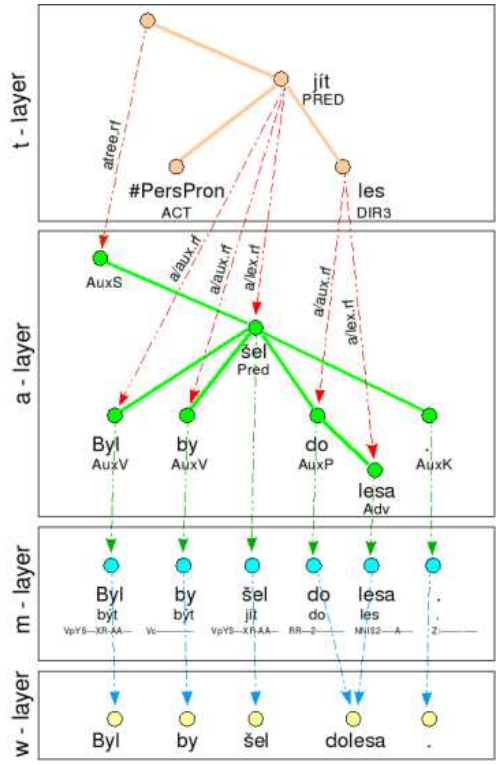


Figure 5.1: Interlinked layers of sentence representation in PDT 2.0 (Sample sentence: *Byl by šel dolesa.* lit. [He] would have gone intoforest.)

5.3 Coordination

So far, we silently assumed that the edges in dependency trees correspond only to dependency relations. However, in the real language there are also non-dependency types of relations between words, and they have to be captured in our representation of the sentence structure too. Probaly the most important representant of such relations is coordination, which is a notorious nightmare for dependency-oriented linguists.[6]

Why does coordination cause difficult problems? There are several reasons. First, the graph the edges of which cover all dependency and coordination relations in a sentence cannot be a tree any more. In the sentence *I saw*

[6]This is not to say that an adequate treatment of coordination is simpler in constituency-oriented approaches. For instance, certain coordination constructions are solved by adding non-tree edges (called *secondary edges*) into the phrase-structure trees in Tiger corpus ([Brants et al., 2002]).

Petr and John there are dependency relations between *Petr* and *saw* and between *John* and *saw*, but there is also a coordination relation between *Petr* and *John*. Thus the graph contains a circle. Second, coordination conjunctions do not fit into the governor/dependent dichotomy, therefore if they should be present in the sentence representation and if the sentence representation should form a connected graph, then the edges incident with the conjuction node do not represent a dependency. Third, dependency relations can be multiplied by coordination relations, e.g. in the expression *Yesterday I saw him and he saw me*, where the temporal modifier is dependent on both verbs. Fourth, a difficult situation arises if coordination and ellipsis are combined: *Yesterday I gave a book to Mary and a doughnut to John*. Fifth, unlike in the case of dependency, more than two units can be related by coordination: *I saw Mary, John, and Petr.*

Insisting on treeness of the formal sentence representation in PDT has two consequences: (1) new special non-dependency types of edges have to be introduced into the dependency tree, (2) some relations intuitively present in the sentence do not have edges of their own in the tree (but are not lost, since they can be reconstructed by composition of other edges).

In this thesis we follow the solution used in the annotation scheme of PDT 2.0.[7] It is based on the following principles:

- Coordination conjunction has its own node in the tree structure.
- The conjunction node is attached under the node which governs the whole coordination construction (the node on which the conjuncts are dependent).
- The conjunct nodes are attached under the coordination node (such edges can be viewed as having a dependency and a coordination component, but both these components have to be composed with some other edge to form a full-fledged dependency or coordination relation).
- If there is a shared modifier dependent on all the conjuncts, it is attached under the coordination node too. There is a special node attribute which distinguishes conjunts (coordination members) from shared modifiers. Coordination members are marked with "M" in the dependency trees in this thesis.

In the dependency trees, it is usual to speak about the nodes which are dependent on other nodes as of their children (with the governing node being their parent). But since not every edge now represents a dependency, we suggest to use the following terminological distinction: *direct children* of N are the nodes which have an incidental edge with the node N and are more

[7] We are aware of the fact that this is not the only treeness-preserving solution: the other solution is e.g. to make the left-most conjunction the head of the construction with the conjunction word and the other conjunct modifying it (as in [Mel'čuk, 1988]).

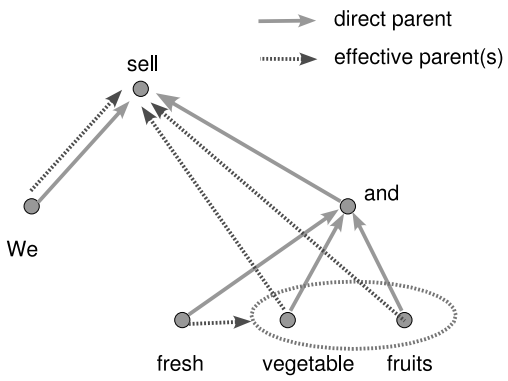


Figure 5.2: The difference between direct parents and effective parents in the PDT 2.0-style trees, illustrated on the sentence *We sell fresh vegetable and fruits.*

distant from the tree root (children according to tree topology), whereas *effective children* of N are those nodes which are linguistically dependent on N. And if N_1 is a direct (resp. effective) child of N_2, then N_2 is its direct (resp. effective) parent. As it follows from the above principles, in the tree topology the effective parent of a node N does not have to be necessarily its direct parent, but also its sibling, grand-father etc. The difference between direct and effective parent is illustrated in Figure 5.2

It is not possible to study valency in the PDT 2.0-style dependency trees without the notion of effective children/parents. The reason is simple: the constraints imposed on nodes by valency frames do not apply to the direct children of the nodes that evokes the valency frame, but rather to its effective children.

Because of the possibility of embeded coordination structures, we suggest to extend also the terminology by introducing the terms *direct coordination member* and *terminal coordination member*. The distinction is characterized in Figure 5.3. Direct coordination member of a coordination structure rooted by node N_0 is a node N_1, which is a direct child of node N_0 and which bears the coordination-member mark. Terminal coordination member of a coordination structure rooted by node N_0 is a node N_x for which it holds that N_x itself is not a root of a coordination structure, and that there is a path $N_0 \ldots N_x$ where for each pair of neighboring nodes N_n and N_{n+1} it holds that N_{n+1} is a direct coordination member of the coordination structure rooted by N_n.

Supposing we have a subtree of a dependency tree, we distinguish the *(direct) root* and the *effective root(s)* of the subtree: if the root of the subtree is a coordination node, then the set of effective parents of the subtree is identical with the set of terminal coordination members. Otherwise the effective root of the subtree is identical with the (direct) root of the subtree.

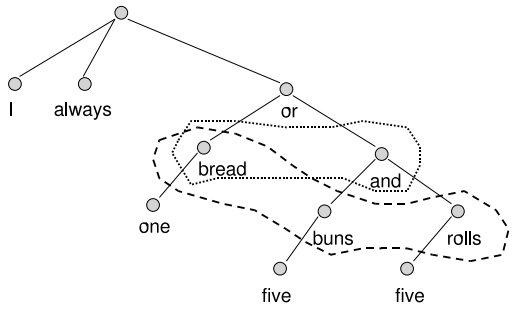


Figure 5.3: The difference between the set of direct coordination members (dotted border) and the set of terminal coordination members (dashed border) illustrated on the disjunction coordination in the simplified dependency tree of the sentence *I always buy one bread or five rolls and five buns.*

The notion of coordination meets with the study of valency also because of several other reasons:

- a criterion called zeugma can be used for distinguishing different LUs ([Stevenson, 2003] page 30: **Arthur and his driving license expired last Thursday*),
- coordination occurs in reciprocal alternation (*Peter met Mary* vs. *Peter and Mary met*),
- if a node has more effective parents, it can fill different slots of their frames (*He came and was immediately recognized by everyone* – the personal pronoun plays the role of actor in the first clause and the role of patient in the second clause)

Remark on tree dimensionality. Unlike in PDT 2.0, coordination does not have a node of its own in the tree structure in FGD, but is represented by a special type of "bracketing". In this context, one can find the idea of a multidimensional dependency tree in the literature related to FGD (e.g. in [Sgall, 1998]) – new dimensions correspond to the coordination brackets. However, in our opinion, it is more advantageous to have a special coordination node and to speak only about different type of partial orderings within the plane dependency trees than to introduce the notion of tree dimensionality. We have the following arguments:

- There is no concept of tree dimensionality in the standard graph theory and the term "tree dimension" does not evoke any natural interpretation.
- From the formal point of view, introducing tree dimensionality is not necessary. We agree that the ordering of direct children of a conjunction bears a different information than the ordering e.g. of actants directly below a verb, but once we differentiate between different types

of partial orderings expressing different language phenomena, the tree may be still "drawn" in a plane.

- Common-sense perception of different types of orderings as different dimensions is problematic as soon as the number of dimensions grows above three or four. When adding new dimensions to capture coordinaton, the number of dimensions can theoretically grow above any limit (see [Sgall, 1998], page 22). Is that really necessary to struggle with such troubles when imagining the additional dimensions in the simple sentence *I met Mary and she told me about Petr and John*?
- The question arises how the nodes of two parallel subtrees should be ordered in the added dimension: *John and Peter met Marry, Sarrah and Lucy* – it makes no sense to say that the node related to *John* has the same position in the third dimension as that of *Mary*, or that the node related to *Peter* precedes that of *Lucy* in this dimension? And if there is no total ordering on the added coordinate, why is that useful to speak about a dimension?

5.4 Two-tiered basic valency model

In this section we introduce a new terminology for describing valency frame instances in surface and deep dependency trees. To our knowledge, no such terminology was ever proposed in the context of FGD (nor in PDT), which makes speaking about instances of valency frames in real syntactic structures quite difficult.

As many authors pointed out ([Helbig and Schenkel, 1969] among the first), the relation between manifestations of valency on the levels of surface and deep syntax (however differently different authors called them) is not straightforward or isomorphic, therefore the two levels deserve their own descriptions of valency. As for naming the two levels, unfortunately we have to face the following terminological Babel:

- "syntactic valency" and "logical-semantic valency", for instance in [Helbig and Schenkel, 1969]
- "valency" and "argument structure", e.g. in [Manning and Sag, 1998]
- "subcategorization" and "valency"[8]
- "valency" and "intention", e.g. in [Daneš and Hlavsa, 1987] (the term "intence" (intention) was introduced by [Pauliny, 1943]),
- "grammatical sentence patters" and "semantic (propositional) sentence patterns", e.g. in [Daneš, 1994]
- "syntactic valency" and "semantic valency", e.g. in project GREG[9]

[8] E.g. http://www.coli.uni-sb.de/~gj/Lectures/DG.LOT/subcat+valency.pdf
[9] http://www.informatik.uni-stuttgart.de/ifi/is/greg-index.html

or in [Karlík, 2000]

- "surface valency" and "deep valency", e.g. in [Pala and Smrž, 2004]
- "C-selection" and "S-selection", e.g. in [Babby, 1998]

In this section, we will use systematically the adjectives "surface" and "deep" for distinguishing concepts from the different syntactic layers. In PDT 2.0 terminology, the former is related to concepts from a-layer and m-layer[10] (because we need both m-layer and a-layer node attributes for saying that something is e.g. a prepositional group in a given case), and the latter corresponds to t-layer.

We will use the following pairs of terms:

- *(deep or surface) valency frame* is a sequence of frame (deep or surface) slots,
- *(deep or surface) frame slot* contains a set of constraints on what can (or must) be filled into this slot,
- *(deep or surface) frame evoker* is a part of the (deep or surface) sentence representation, which represents the frame-evoking lexical unit,
- *(deep or surface) frame slot filler*[11] is a part of the (deep or surface) sentence representation, which "saturates" one of the frame slots of the frame evoked by the frame evoker,
- *(deep or surface) frame instance* is a part of the (deep or surface) sentence representation, in which the frame usage is manifested; frame instance consists of a frame evoker and frame slot fillers.

5.4.1 Surface, deep, and complex valency frames

Surface valency frame is a sequence of slots where each slot contains some constraints on what types of sentence elements can be filled into this slot: for instance the slot has to be filled with a noun in dative, with a certain prepositional group, or with a certain type of subordinated clause. These constraints are described in more detail in Section 5.6.

[10] In the recent versions of FGD (since 1990's), surface syntax is not treated as an autonomous level of language description. However, in our opinion the surface-syntactic structures will be indispensable (not only because of technical, but also because of theoretical reasons) when attempting at a fully formal description of valency. For instance, when disregarding this intermediate layer, we do not have any means to express that a given actant of a given verb can be expressed by a subordinating clause introduced by a specific conjunction. Subordinating clause as a syntactic construct can be hardly seen just as a morphemic form of the verb heading the clause, neither is the notion of subordinating conjunction present on the tectogrammatical level.

[11] Also other authors have pointed out that it is necessary to make the distinction between what we call here slot and slot filler. [Mel'čuk, 2004] distinguishes between actant slots and actants: "Informally, an actant slot of L in the lexicon is and "empty place" or "open position" foreseen in the lexicographic description of L ... each of L's slots has to be "filled" or "saturated" with a linguistic entity of a particular type".

Deep valency frame is a sequence of slots where each slot contains (at least) two types of constraints: what is the functor of the sentence element filling this slot (5.6), and whether this filling element is obligatorily present in the deep-syntactic tree.

If the surface and deep valency frames of the same lexical unit are alligned in parallel, the result can be thought as a two-tiered table which we call *complex valency frame*.[12] In valency lexicon, each LU is supposed to be associated with one complex valency frame.

5.4.2 Surface and deep frame evokers

Surface frame evoker (SFE) is a subgraph of the a-layer dependency tree having a valency potential (evoking a valency frame). In case of verbal lexical units, a surface frame evoker covers all nodes representing the (possibly complex, possibly incomplete or completely deleted) verb form expressing the given LU. Unlike SFE, *deep frame evoker* (DFE; frame-evoking subgraph of t-layer tree) is a single t-layer node in most cases, since complex verb forms are reduced to one t-layer node, and deleted autosemantic nodes are restored on the other hand.

In the simplest case, both SFE and DFE are formed by one node and thus there is a trivial one-to-one correspondence between SFE and DFE nodes, as shown in Figure 5.4 (a). But also the following non-trivial situations (and their combinations) have to be considered:

- SFE represents a reflexivum tantum and thus contains a reflexive particle, Figure 5.4 (b),
- SFE corresponds to a complex verb form and contains an auxiliary verb (or verbs), Figure 5.4 (c),
- deleted verb form, Figure 5.4 (d),
- incomplete complex verb form, Figure 5.4 (e),
- two overlapping evokers – shared auxiliary verb, Figure 5.4 (f),
- two overlapping evokers – haplology, Figure 5.4 (g),

5.4.3 Surface and deep frame slot fillers

Surface slot filler (SSF) is a subgraph of an a-layer dependency tree which saturates the valency potential of the governing lexical unit, realized as SFE

[12] The reader immediately reveals that our termonilogy in this point is inspiration by [Daneš and Hlavsa, 1987]. However, we deliberately do not adopt their terms (complex sentence pattern = grammatical sentence pattern + semantic sentence pattern), since we work in a significantly different framework (especially when considering deep syntax), and thus using their terms would lead to a confusion.

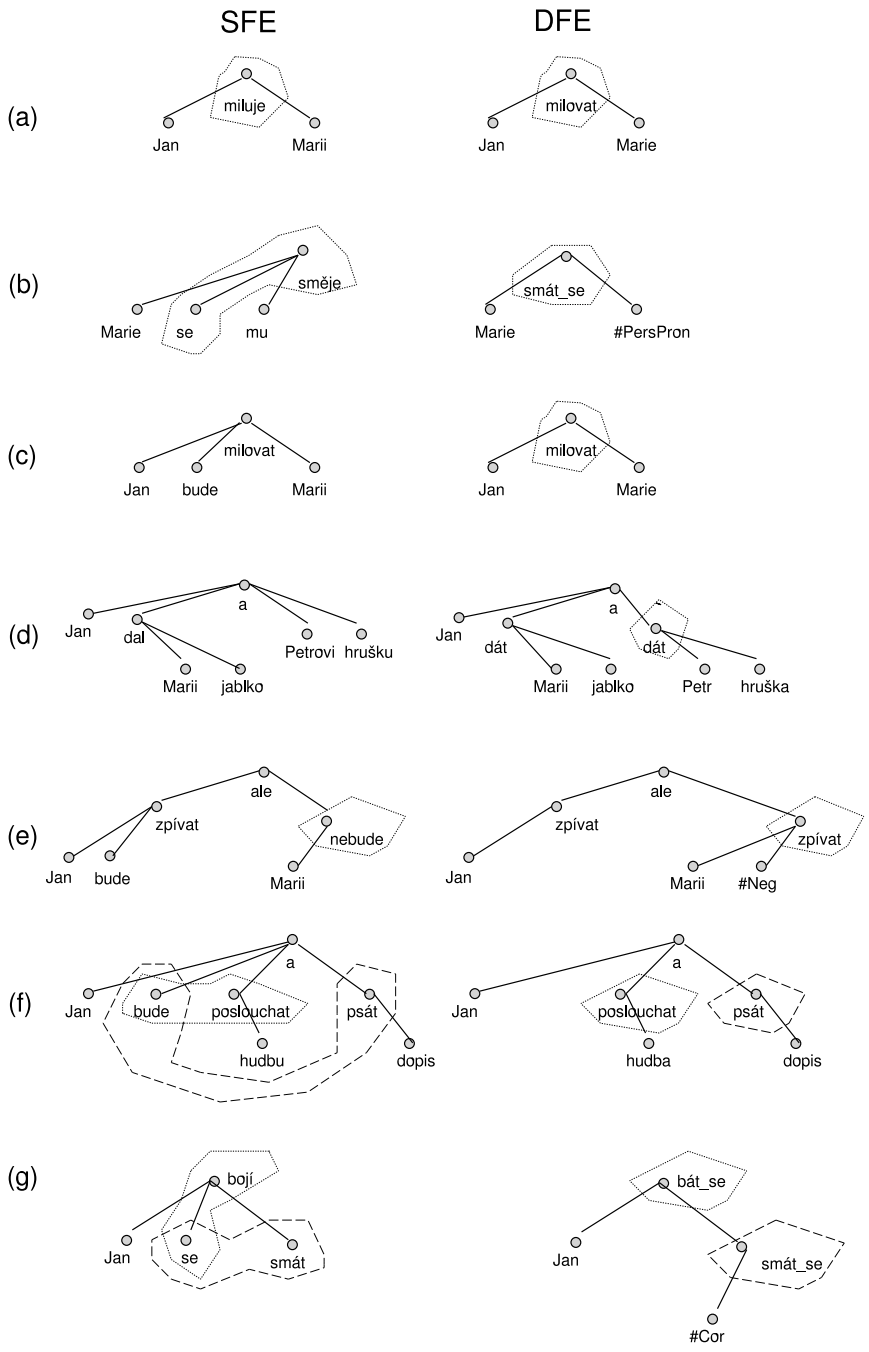


Figure 5.4: Deep and surface frame evokers in (simplified) t-layer and a-layer trees: (a) *Jan miluje Marii* (lit. John loves Mary) (b) *Marie se mu směje* (lit. Mary refl-particle at-him laughts) (c) *Jan bude milovat Marii* (lit. John will love Mary) (d) *Jan dal Marii jablko a Petrovi hrušku* (lit. John gave Mary apple and Petr pear) (e) *Jan bude zpívat, ale Marie nebude* (lit. John will sing, but Mary won't) (f) *Jan bude poslouchat hudbu a psát dopis* (lit. John will listen to-music and write letter) (g) *Jan se bojí smát* (lit. John refl-particle is-frightened to-laught)

in the same tree. By analogy, *deep slot filler* (DSF) is a subgraph of a t-layer dependency tree which saturates the valency potential of the governing lexical unit, realized as DFE in the same tree.

In the simplest case (infinitives or prepositionless morphological cases), SSF is formed by one single node, as in Figure 5.5 (a). Coordination can come into play, as illustrated by Figure 5.5 (b).

In the case of prepositional cases or subordinated clauses (starting with a subordinating conjunction), the SSFs contain also the functional words, as illustrated by Figures 5.5 (c) and (d) (if the verb in the SSF subordinated clause had a complex verb form, then also the remaining parts of the complex verb form should be included into the SSF). As the coordination can be located "above" or "below" the functional word, the number of possible combinations grows, see Figures 5.5 (e) and (f).

In the following list, some more intricate cases of the SSF and DSF relation will be mentioned:

- DSF has no a-layer counterpart and SSF is empty. There are two possible reasons:
  - Corresponding SSF cannot be expressed on the surface at all because of grammar rules, e.g. the subject of infinitives. In this case, the ficitious lexeme #Cor is used in the restored t-layer node, see Figure 5.6 (a).
  - The corresponding SSF could be expressed, but was deleted from the surface shape of the given sentence, e.g. because of pro-drop, see Figure 5.6 (b) (DSF is a restored node).
- DSF has no a-layer counterpart, but SSF is not empty. This happens e.g. when a noun is elided in the surface shape of the sentence and only an adjective attribute – dependent on a virtual noun – is present. Thus SSF contains only the adjective, while DSF contains the restored node with a ficitious lexeme. See Figure 5.6 (c).
- SSF contains a demonstrative pronoun in an expletive position,[13] while DSF corresponds to the verbal head of the subordinated clause dependent on the pronoun (but in PDT 2.0, SSF is still included among the a-layer nodes interlinked with the t-layer verb). See Figure 5.6 (d).
- SSF and DSF may correspond to two completely (non-empty) disjunct parts of the original sentence due to the fact that a-layer and t-layer dependencies may have reverse directions in certain situations. For instance, Czech numerals higher than four govern the counted nouns on

[13] Expletive is a syllable, word, or phrase inserted to fill a vacancy (as in a sentence or a metrical line) without adding to the sense; especially : a word (as in "make it clear which you prefer") that occupies the position of the subject or object of a verb in normal English word order and anticipates a subsequent word or phrase that supplies the needed meaningful content (http://www.webster.com)

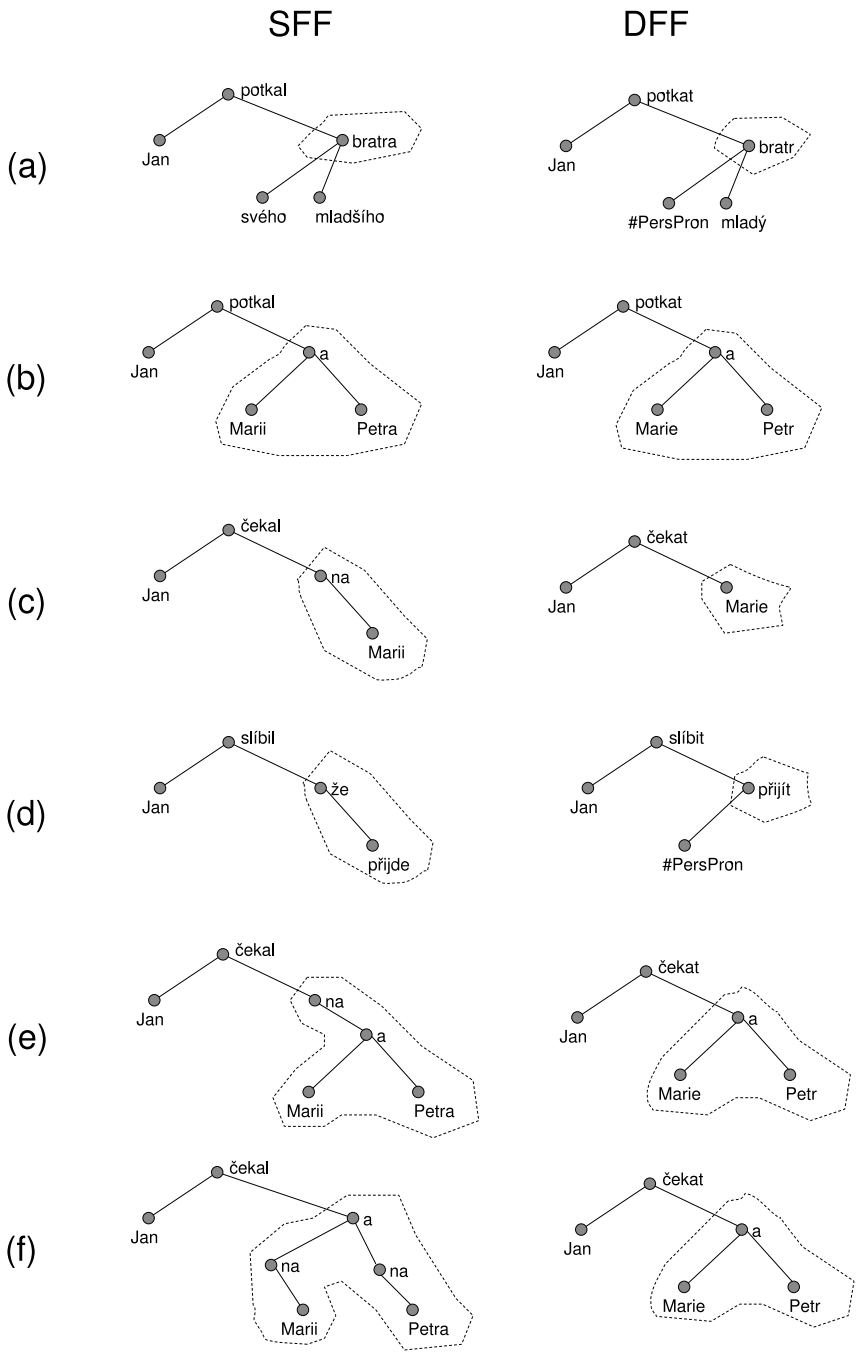


Figure 5.5: Examples of pairs of deep and surface frame fillers in (simplified) t-layer and a-layer trees: (a) *Jan potkal svého mladšího bratra* (John met his younger brother) (b) *Jan potkal Marii a Petra* (John met Mary and Peter) (c) *Jan čekal na Marii* (John waiter for Mary) (d) *Jan slíbil, že přijde* (lit. John promised that he-will-come) (e) *Jan čekal na Marii a Petra* (lit. John waited for Mary and Peter) (f) *Jan čekal na Marii a na Petra* (lit. John waited for Mary and for Petr)

a-layer (because of the morphological agreement), whereas on t-layer the numerals are dependent on the counted nouns. Thus in these cases SSFs contain the numerals, while DSFs contain the counted nouns, as depicted in Figure 5.6 (d).

5.5 Constraints on surface frame evokers

In most cases, all (conjugated) lexical forms of a given lexeme can be associated with all lexical units of the lexeme. However, there are exceptions where the usage of certain lexical forms of the lexeme is not compatible with some of its lexical units. Or in other words, usage of a certain lexical unit may impose special constraints on the form of the surface frame evoker. Examples:

- Aspect. The most frequent case is the constraint on the morphological aspect, where either only the perfective or only the imperfective forms can be used:

  (13) Výrobek odpovídal očekávání.
  The product fulfilled the expectations.

  (14) *Výrobek odpověděl očekávání.
  *The product answered their expectations.

- Passive. Some lexical units require the frame evoker to be passivized:

  (15) Je připraven jít tam.
  He is ready to go there.

  (16) *Připravili ho jít tam.
  They made him ready to go there.

  (17) Byli odkázáni na cizí pomoc.
  They were entirely dependent on the outer help.

  (18) *Odkázali je na cizí pomoc.
  They made them dependent on the outer help.

- Negation. Negation can be required or forbidden for a given LU:

  (19) Nedá mu to tam nejít.
  He is tempted to go there.

  (20) *Dá mu to tam nejít.
  He is not tempted to go there.

- Tense

  (21) Dám na to krk.
  I bet my head.

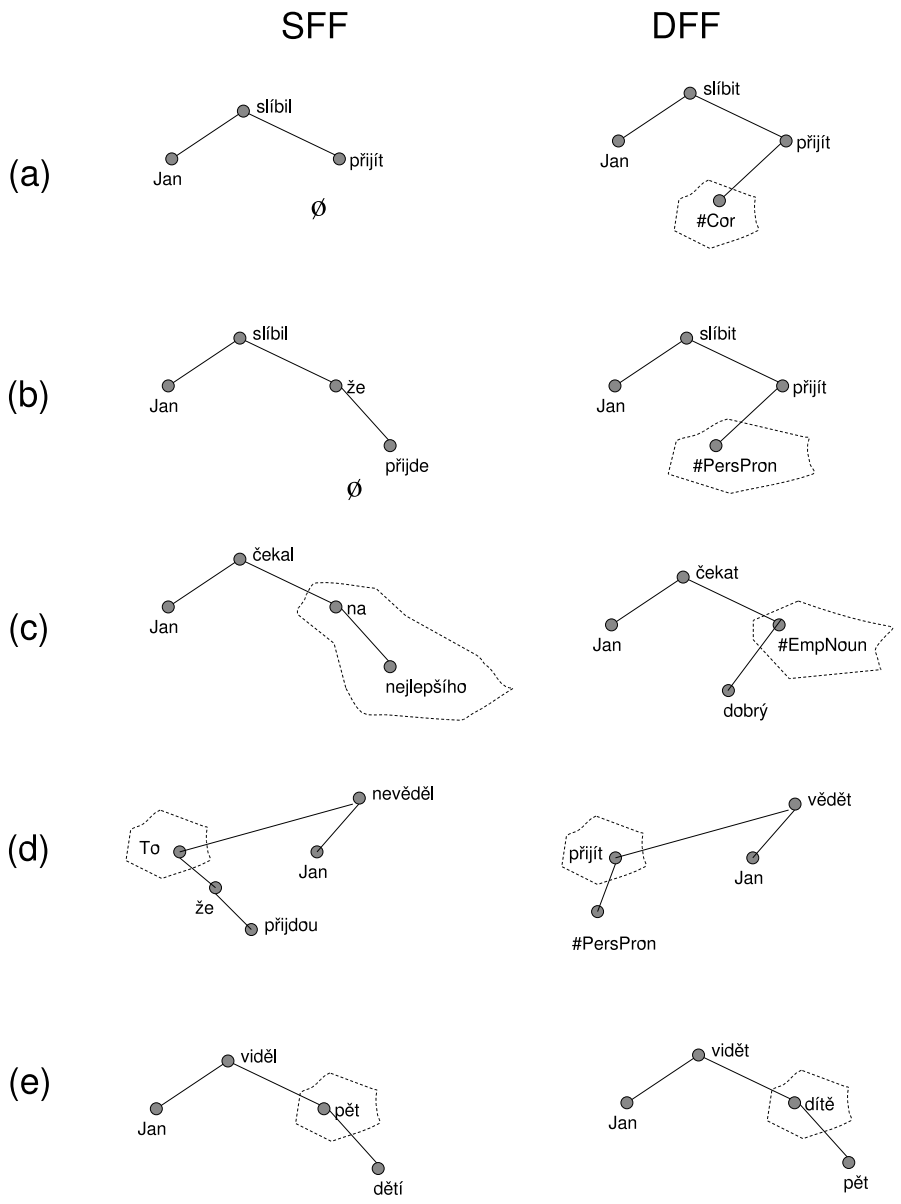


Figure 5.6: Examples of non-correspondences between deep and surface frame fillers in (simplified) t-layer and a-layer trees: (a) *Petr slíbil přijít* (lit. He promised to-come) (b) *Petr slíbil, že přijde* (lit. Petr promised that he-comes) (c) *Jan čekal na nejlepšího* (lit. John waited for the-best) (d) *To, že přijdou, Marie věděla.* (lit. It that they-come Mary knew) (e) *Jan viděl pět dětí* (John saw five children)

(22) *Dal jsem na to krk.
I bet (past) my head.

More examples can be found in [SČFI, 1983] (the potentially constrained features are listed also in [Hnátková, 2002]).

The information about constraints on frame evoker form would be useful especially for word sense disambiguation. However, these constraints have not been annotated in the presented lexicon yet.

5.6 Constraints on surface frame slot fillers

In the surface shape of a sentence, the set of morphemic means that can be used for expressing a particular frame slot filler of a particular frame evoker is always limited. The reason is simple: if there was no limitation, one could hardly recognize which expression in the sentence fills which slot of the present frame evoker.

When explicitly describing constraints of surface realizations, it is necessary to find an adequate level of generalization: obviously it is too general to say that the patient of the verb "to wait" must be a prepositional group (as "She waited into Peter" is not grammatical), and it is too specific to say that it must be preposition "for" with a noun (as the noun can be substituted with a personal pronoun).[14]

We distinguish two types of constraints on the surface realizations:

- *Explicit constraint* – one or more possible surface realizations are stored in the lexicon for the given slot. Surface frame slot filler should match one of them.
- *Implicit constraint* – the set of possible surface realizations of a given slot is implied solely by the functor of the slot; in other words, any surface form expressing the given functor is acceptable. Therefore no constraints have to be explicitely stored in the lexicon in this case. It happens especially in free-modifier slots.

Such constraints are lexically specific,[15] that is why they should be contained in the lexicon.

In VALLEX, we cluster the explicit constraints imposed on frame slot fillers into the following classes:

Explicit constraints used in VALLEX can be clustered into the following classes:

[14] This is a trivial observation, but from the implementational point of view it is not always easy to merge nouns and "noun-like" pronouns without overgeneralizing, as many morphological tagsets do not support the distinction beetween nominal and adjectival pronouns.

[15] Although there is often a strong tendency to use the prototypical form for a given functor (e.g. dative for addressee in Czech), one can always find exceptions.

- *Pure (prepositionless) case.* There are seven morphological cases in Czech. In VALLEX, we use the traditional numbering to denote the case: 1 - nominative, 2 - genitive, 3 - dative, 4 - accusative, 5 - vocative, 6 - locative, and 7 - instrumental.
  SSF matches this constraint if its effective root is a noun or a nominal pronoun in the given case. In the case of surface deletion of the noun, the SSF contains an adjective (or an adjectival pronoun) instead, e.g.:

  (23) Petr žádného neviděl.
  Petr saw none.

- *Prepositional case.* This constraint specifies lemma of the preposition (i.e., preposition without vocalization) and the number of the required morphological case (e.g., *z+2*, *na+4*, *o+6*...). The following prepositions occur in VALLEX: *bez, do, jako, k, kolem, kvůli, mezi, místo, na, nad, na úkor, o, od, ohledně, okolo, oproti, po, pod, podle, pro, proti, před, přes, při, s, u, v, ve prospěch, vůči, v zájmu, z, za.*
  SSF matches this constraint if its effective root is the given preposition and the effective child of the preposition is a noun or a nominal pronoun in the given case. In the case of surface deletion of the noun, the SSF contains only an adjective or an adjectival instead in the given case, e.g.:

  (24) Petr na žádného nečekal.
  Petr waited for none.

- *Subordinated clause.* Lemma of the conjunction is specified. The following subordinating conjunctions occur in VALLEX: *aby*, *ať*, *až*, *jak*, *zda*, *že*. Note: the form *zda* is in fact an abbreviation for the couple of conjunctions *zda* and *jestli*.
  SSF matches this constraint if its effective root is the given subordinating conjunction, and its effective child is a verb (head of the subordinated clause).
  In rare cases, the head of the subordinated clause is supposed to be negated:

  (25) Nedá mu to, aby tam nešel.
  He is tempted to go there.

  but

  (26) *Nedá mu to, aby tam šel.
  (untranslatable)

  Ideally, this requirement should be captured in the lexicon too.

- *Infinitive.* The abbreviation 'inf' stands for infinitive verbal complementation.
  SSF matches this constraint if its effective root is an infinitive.

- *Constructions with 'být'.* Infinitive of the verb *být* (to be) may combine with some of the types above, e.g. být+adj-1, e.g.:

  (27) Zdá se to být dostatečné.
  It seems to be sufficient.

  SSF matches this constraint if its effective root is the verb *být*, the effective child of which matches the other half of the constraint.

- *Construction with adjectives.* The abbreviation 'adj-digit' stands for an adjective (or an adjectival pronoun) in the given case, e.g. adj-1 for adjective in nominative.
  Ideally, the lexicon should also contain the information about agreement: it should be specified with which from the remaining slot filler the adjective should agree in gender and number. For instance, the adjective in the frame of the verb *cítit se* (to feel) agrees with the actor (*Marie se cítí slabá* – Mary feels weak$_{fem}$), whereas in the case of 'považovat' (to find) it agrees with patient (*On ji považuje za chytrou* – He finds her clever$_{fem}$).[16]
  SSF matches this constraint if its effective root is an adjective (or an adjectival pronoun) in the given case.

- *Content clause.*[17] Abbreviation 'rel' is used in VALLEX. Example:

  (28) Neví, kolik to stojí.
  He does not know how much it is.

  SSF matches this constraint if its effective root is the head of a content clause.

- *Direct speech.* Direct speech can be introduced as an actant of a verb, especially in the case of verbs of speaking.[18]
  SSF matches this constraint if its root or its effective root is the root of a direct speech subtree.[19]

- *Dependent part of phraseme.* If the set of the possible lexical values of the given complementation is very small, we list these values directly (e.g. *napospas* for the phraseme *ponechat napospas* (to expose), or *z kopýtka* for phraseme *vyhodit si z kopýtka* (to spree)).[20]

[16] Note that the agreement really concerns the patient, not the object: *Ona je jím považována za chytrou* – she is considered as clever by him.$_{fem}$

[17] *Vedlejší věta obsahová* in Czech.

[18] Note that direct speech can naturally occur below many other verbs (e.g. verbs of emotion) in Czech – unlike English and similarly to Russian, see [Mel'čuk, 1988]. page 341. Mel'cuk's example: *"Ostav'te menja!" – ispugalsja bufetčik* (*"Leave me alone!" – said the bartender, frightened* – the verb 'to say' had to be added). Note that in such cases the direct speech does not fill any frame slot.

[19] This is difficult to recognize in the surface-syntax tree, since direct speech can contain really anything.

[20] In case of multiword parts of phrasemes, the tree (and not only the sequence of

SSF matches this constraint if the sequence of the forms of its nodes equals (string-wise) the prescribed form.

When testing the above constraints on the surface-syntactic trees of real sentences, one can meet numerous violations caused by usage of named entities:

(29) Proti všem jsem ještě nečetl.
I have not read Proti všem (Against all) yet.

or by meta-usage of words of various part of speech ([Grepl and Karlík, 1986], page 237:

(30) Každé proč má svoje proto.
Every why has its that's why.

5.7 Functors, subfunctors, and superfunctors

In the FGD terminology, the term functor[21] labels the relation between a tectogrammatical unit and its governor. It denotes what is called semantic role, theta role or deep syntactic relation in other approaches.

Some authors pointed out that there is surprisingly little consensus on what the set of linguistically significant role types is, and that "a small fixed set of thematic roles has never been agreed on, and it seems unlikely that this will change" ([Davis, 2001], page 20). However, certain regularities can be observed if we mentally order the existing systems along the scale from *ordered argument systems* to *true thematic role systems*. This distinction was suggested in [Dowty, 1986]. An ordered argument system is simply a scheme for distinguishing between the arguments of a verb. Example of this approach is PropBank, where the arguments are only numbered and no common properties associated with the given argument number are guaranteed. In the true thematic role systems, each thematic role implies some properties and allows for some interpretation of the given argument. The system of functors as developed in FGD is located somewhere in between: on one hand, the shifting principle (mentioned in Section 2.1.2) causes that no generally valid properties can be stated for the cognitive counterparts of ACT and PAT. This holds to certain degree also for EFF, which is often used just as 'the third' actant, without having to do with the real effect of the

forms) representing this part should be ideally captured in the lexicon (the noun *kopýtko* dependent on the preposition *z*).

[21] Remark on terminology: the term of functor as introduced in FGD and used in this thesis has nothing to do with the purely mathematical notion of functor in category theory. In category theory, a functor is a mapping from one category to another which maps objects to objects and morphisms to morphisms in such a manner that the composition of morphisms and the identities are preserved. (http://encyclopedia.thefreedictionary.com).

process expressed by the verb, and thus EFF looses its semantic homogenity too.

In FGD, the functors were originally used only for capturing the type of dependency relation between a tectogrammatical node and its governor. As it was already mentioned in the Section 2.1.2, the original (dependency) functors are divided into *actants* and *free modifiers*.[22] Later, the inventory of functors was enriched during the development of the PDT – the tectogrammatical nodes can be labeled also with new functors having not to do with dependency. There are functors serving for distinguishing the types of the very root of the tectogrammatical tree (PRED for predicate head vs. DENOM for nominal head), functors for several types of coordination (CONJ, DISJ, ADVS, etc.) and apposition (APPS), functors for marking parenthetical constructions, a functor for marking expressions in foreign languages (FPHR), functors for marking a dependent part of a phraseme (DPHR) and a nominal part of verbonominal predicate (CPHR), etc. Thus the functor attribute of t-layer nodes is highly polyfunctional in PDT 2.0.

The complete list of functors used presently in VALLEX is attached in Appendix A. However, the system of functors is still in development.

In the conventional FGD valency frames, each frame slot is associated with one functor. The question arises whether it is sufficient. The system of functors is based on certain generalizations,[23] and we are not aware of any common-sense reason why different verbs should not require different degrees of generalization for capturing their valency potential. It might be useful to have a hierarchy of functors, with some of them subsuming the others, instead of having a set of isolated atomic values.[24]

When creating the hierarchy, we can structure the set of functors in two directions: to higher specificity, and to higher generality.

More specific functors, called *subfunctors*, are already used in PDT 2.0, thus forming a two-level system together with functors. For certain functors, a subfunctor can be used to state the semantic roles between the dependant and the governor more precisely. For instance, functor DIR3 can be combined with 12 subfunctors: DIR3.betw (*Spadl mezi stoly* – He fell between the tables), DIR3.above (*Spadl nad stůl* – He fell above the table), DIR3.below (*Spadl pod stůl* – He fell below the table), DIR3.front (*Spadl před stůl* – He fell in front of the table.), DIR3.behind (*Spadl za stůl* – He fell behind the table.), DIR3.elsew (*Spadl mimo stůl* – He fell outside the

[22]Note that this dichotomy is named in many different ways in the literature (see the discussion about the prefered term of actant in [Mel'čuk, 2004], page 6): participants vs. circumstantials, inner/internal participants vs. outer/external participants, arguments vs. adjuncts, complement vs. modifier, etc.

[23]Unlike the constraints of surface forms, functors are of course not observable in a sentence 'by the naked eye', and thus require higher degree of abstraction. Thus the danger of being trapped in a labyrinth built on arbitrary decisions is also much higher.

[24]The question of atomicity was mentioned in the critique of the traiditional view of "thematic roles" in [Davis, 2001], page 20.

table.), etc. From the given examples, it seems that all subfunctors are applicable with the given verb. One can observe that if a certain prefixed verb[25] is used, certain prepositions (and therefore also certain subfunctors) are less likely to be used in comparison to others: for instance, the sentence *Zapadl před stůl* (lit. He-fell-behind in-front-of table.) does not sound natural, but is meaningful and grammatically correct. These subfunctor preferences are not included in VALLEX at the present stage.

The issue of more general functors – let us call them tentatively *superfunctors* – seems to be much more important in the study of valency. In the context of FGD, a notion similar to superfunctors has been alluded already several times: for the first time probably in [Panevová, 1980] page 52 ("hierarchization"), recently also in [Urešová, 2004] ("groups of functors") or in [Součková, 2005], page 40 ("hypermodification"). Such means are useful for describing the situations where expressions with two (or more) different functors seem to fill the same valency slot. The hypothesis of superfunctors is supported by at least two types of observations:

- an expression having a certain functor can be regularly substituted by a group of expressions with other functors:

  (31) Odložil to o dva dny.DIFF.
  He postponed it by two days.

  (32) Odložil to o z pondělka.TFRWH na středu.TTIL.
  He postponed it from Monday to Wednesday.

- two expressions with different functors can be coordinated:

  (33) Telefonoval jsem do jeho školy.DIR3 i jeho rodičům.ADDR.
  I called both to his school and to his parents.

At the moment, we preliminarly suggest the following superfunctors (however, they are not captured in the present version of VALLEX yet):

- G_ADDR_DIR3 (generalized addressee) - can be saturated by ADDR or DIR3 or both,
- G_ORIG_DIR1 (generalized origin) - ORIG or DIR1 or both (symmetrically to G_ADDR_DIR3),
- G_T_DIFF (generalized temporal difference) - DIFF or TFRWH+TTIL or DIFF+TFRWH or DIFF+TTIL,
- G_DIR (generalized direction) - DIR1 or DIR2 or DIR3 or any combination.

The same interchangability principle as in the case of superfunctors can be applied also on some basic functors, without introducing new superfunctors. For instance, THL can be substituted by TFRWH and TTIL.

[25] The relation between verbal prefixes and prepositions was studied in [Bémová, 1979].

Note that the superfunctors cannot simply replace actants or free modifiers in all situations: *zaměřit něco někam* (to aim something somewhere) cannot be combined with addressee, whereas in *vyprávět něco někomu* (to tell someone something) the addressee cannot be substituted with a direcitonal modifier.

5.8 Selectional preferences

Some verbs prefer arguments of a particular semantic type. For instance, the patient of the verb "to eat" is usually something than can be eaten, and the actor of "to bark" is usually a dog. Originally, the term "selectional restrictions" (meant to be hard constraints) was used ([Katz and Fodor, 1963], [Chomsky, 1965]),[26] for these regularities. But many researchers soon pointed out that such constraints should be understood as preferences rather than as restrictions in any strict sense. Let us cite from [Sgall et al., 1986], page 106:

> A sentence meeting the restrictions of strict subcategorization, but not the selectional restrictions, seems to have a meaning (or even several): wittness the fact that it can be translated into other languages[27] ... Thus, it is not advisable to 'asterisk away' sentences that merely haven't found any occasion of use, thanks to our image of the world.

A similar note can be found also in [Mel'čuk, 1988]:

> [Meaning-Text theory] does not distinguish "normal" meanings from absurdities, contradictions or trivialities. Discovering that something is stupid or absurd or detecting contradictions is by no means a linguistic task.

Even if the theoretical status of selectional preferences is problematic, from the NLP viewpoint they represent very important information, especially in the Word Sense Disambiguation task: for instance, if the verb "to expire" is used with non-animate abstract subjects such as *time / deadline / lincense / offer expired*, then it is highly improbable that the verb is used in the sense "to die" or "to breath out". In some situations, selectional

[26] Chomsky's sentence "Colorless green ideas sleep furiously" is probably the most well-know example of violating such restrictions, but not the first one: "The pioneering (yet mostly forgotten) French syntactician Lucien Tesniere came up with the French sentence *le silence vertébral indispose la voile licite* ('the vertebral silence indisposes the licit sail') to make essentially the same point in his 1954 book (which was mostly written in the late 1930s)." http://www.brainyencyclopedia.com

[27] Just for curiosity: fully grammatical translations of the sentence "I can eat glass and it does not hurt me" into lots of languages can be found at http://www.everything2.com/index.pl?node_id=493597

preferences are even the only source of information leading to the correct reading of a sentence (or, better to say, the more likely reading, with respect to the nature of the real world). Czech examples:

- *Lvíče snědlo dítě* (lit. Lionet ate child) – ambiguous because of the fact that both nouns have the same form in nominative and accusative and word order is not sufficient for determing who ate what.
- *Učil matematiku malé školáky* (lit. He-taught mathematics little pupils) – ambiguous because of the fact that both objects of the verb *učit* can be expressed by accusative at the same time and word order is not sufficient to distinguish between what was taught and who was taught.

Roughly since [Resnik, 1993], most experiments related to selectional preferences are based on a combination of pre-defined semantic class hierarchy (most frequently WordNet-based ontologies) with statistical tools.

We do not include the manual annotation of selectional preferences into the annotation scheme of VALLEX. As a lot of work has been done on automatic extraction of selectional preferences for given lexical units, we hope that VALLEX can be enriched with this type of information later in a more or less automatic fashion.

In [Hlaváčková and Horák, 2005] it is suggested to use selectional preferences instead of functors in the description of valency frames. A two-level system of semantic role labels was created, where the first level distinguishes between concepts from EuroWordNet Top Ontology ([Vossen, 1998]), whereas the second level uses literals from Princeton WordNet Base Concepts. We agree that adding such information is important for increasing Word Sense Disambiguation success rate. However, in our opinion the information represented by functors and the information represented by selectional preferences are of completely different nature and none of them can be used as the substitute of the other. Selectional preferences specify which lexical units are more likely to appear as slot fillers of a given lexical unit, whereas functors capture the syntagmatic relation between the slot filler and its governor. Thus almost all nouns from various ontological categories receive the functor PAT when filled into the sentence *'I haven't heard about . . . '*, and vice versa, entities from the same ontological category (or even the same entity) can be labeled with many different functors in different contexts (*Peter came*, *I saw Peter*, *It was Peter's book*, *She came before Peter*. . .). Moreover, even absurd sentences violating the selectional preferences (such as 'Doughnat ate Homer') can be analyzed and labeled with functors.

5.9 Verbs of control

The notion of control (originally developed in Chomsky's framework of Government and Binding) was introduced into FGD in [Panevová, 1996] and

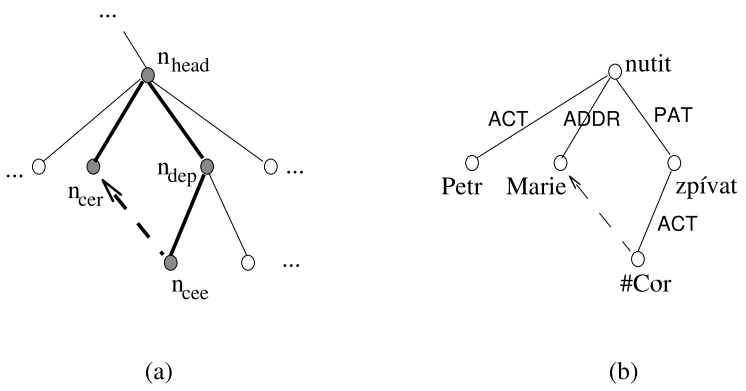


Figure 5.7: Simplified fragments of tectogrammatical trees. (a) general control pattern (b) Control in sentence "Petr nutil Marii zpivat" (the ficitious lexeme #Cor is used for the restored subject in the PDT 2.0).

further elaborated in [Skoumalová, 2001], pages 49-65. Control (as a subtype of grammatical coreference) was annotated in the Prague Dependency Treebank 2.0 ([Kučová et al., 2003]). We only sketch the basic principles (and the consequences for VALLEX) in the following paragraphs.

The generall idea is depicted in Figure 5.7 and can be formulated as follows: let us suppose there are two nodes n_{head} and n_{dep} in a tectogrammatical tree, the latter being dependent on the former. Then there might be a virtual dependant below n_{dep}, called controlee (node n_{cee}), which is referentially identical (coreferential) with one of n_{head}'s children, but which was not expressed on the surface, either because it is not possible due to grammar rules (virtual subjects below infinitives), or as a result of optional deletion (possessives below nominalizations).

Control is lexically conditioned (and therefore lexicographically relevant) because of the fact that the lexical value of Nhead often implies which of the n_{dep} children and which of the Nhead children enter into coreference relation.

In VALLEX, we deal so far only with one specific type of control, in which both n_{head} and n_{dep} correspond to verbs and n_{dep} is realized as an infinitive on the surface. We follow [Panevová, 1996] (note 3) in that "[Controller] is understood here as one of the participants of the governing (head) verb. [Controlee] is always the subject of the dependent verb."

Information about the type of control should be stored in each LU which has infinitive as the surface form in one of its frame slots.

Finally, we would like to note that introducing a restored node with a ficitious lexeme and marking the coreference relation is not the only possible solution how to capture control in dependency trees. The other option used in PropBank ([Kingsbury and Palmer, 2003]) is to draw a direct cross-tree

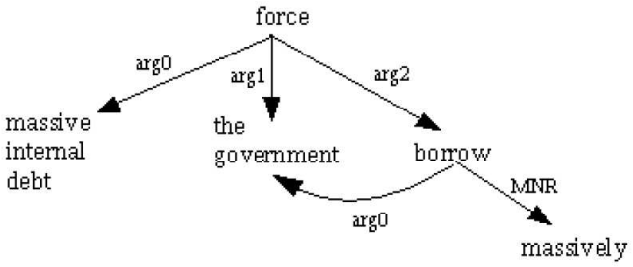


Figure 5.8: Control in PropBank: tree fragment corresponding to "...massive internal debt has forced government to borrow massively ..." (adapted from [Kingsbury and Palmer, 2003]).

dependency edge, as it is depicted in Figure 5.8. But this does not change the way it should be captured in the lexicon.

5.10 Remark on modal verbs

Following [Panevová et al., 1971], the current version of FGD (as well as its implementation in PDT) understands a verb expression containing a modal verb as one tectogrammatical unit. The modal verb is understood as a grammatical means and thus it is represented only as a grammateme value,[28] which is attached to the node representing the meaninful verb. In this section we would like to point out three types of problematic consequences of such an approach.[29]

First, the set of modal verbs does not have sharp borders (or, better to say, it seems to have more blurred border than other sets of functional words, e.g. prepositions).[30] There is a scale from undoubtedly modal verbs such as *muset* (to must), through "less modal" verbs (with increased portion of lexical meaning) such as *hodlat* (to intend), to verbs which are fully meaningful but still signal some modality, e.g. *zamýšlet* (to think of).

Second, modal verbs in Czech do not have the usual property of grammatical categories stated in [Bybee, 1985], page 191:

> The basic idea [of the notion of grammatical category] is that sets of conceptually-related morphemes *contrast* with one another, in the sense that the presence of one excludes the presence of another.

[28] See [Razímová and Žabokrtský, 2005] for the description of the system of grammatemes in PDT 2.0.

[29] Some more items for the discussion can be found in [Panevová et al., 1971], pages 111-118, or [Sgall et al., 1986], pages 170-171.

[30] Problematic delimitation of the set of modal verbs led to cyclic changes during the development of the PDT.

In many cases, the usage of one modal verb in a complex verb expression does not exclude the possibility of usage of other modal verb. Leaving aside coordination constructions, modal verbs can be combined together also hypotactically. 327 instances of such expressions (containing two modal verbs and one meaningful verb) can be found in ČNK by the following simple query:

```
[lemma="(muset|moci|chtít|smět|umět|dovést)"] ([tag="Vf.+" & lemma="
(muset|moci|chtít|smět|umět|dovést)"])[tag="Vf.+"]
```

There are two possibilities how to arrange the tectogrammatical tree so that these expressions can be properly represented: either the internal structure of the node would have to be made significantly more complex (since grammatemes with their atomic values are not sufficient for capturing two modal verbs at once), or at least one of the two modal verbs would be represented as a separate node. In PDT 2.0, the second solution was used. However, both these solutions are highly uncomfortable: the former leads to a formally much more complicated system, whereas the latter requires different treatment of the same modal verbs in different contexts, which is also undesirable.

Third, a modal verb can have its own modifiers, depending specifically on the modal verb, not on the meaningful one. If the two verbs are collapsed into one tectogrammatical node, then there is no way to capture the difference in attachment of adverbs *always* and *once* in the sentence *I always wanted to meet him once*. In the real data, it happens especially in the case of negation, which is represented as a separate tectogrammatical node in the case of verbs. Obviously, the only solution here is to separate the verbal expression into two nodes in such situations. So again, a modal verb is represented once as a separate node and once as a grammateme in the PDT 2.0, depending on whether there is a negation or not.

Taking into account the above observations, we are convinced that it would be more adequate to treat Czech modal verbs as autosemantic verbs, i.e. to represent them as full-fledged tectogrammatical nodes, similarly to verbs of control. The consequence for VALLEX is that modal verbs should be represented in the lexicon as any other verb of control.

5.11 Alternations

5.11.1 Basic and derived lexical units

The basic lexicon model sketched on page 35 relies on two assumptions: (1) the description of a lexeme can be divided into a discrete set of more or less unrelated lexical units, each of them having its own static entry in the lexicon, (2) the syntactic behavior of one lexical unit can be described by one single (two-tiered) valency frame.

The problem is that many verbs can be used in different contexts in slightly different meanings, which can be possibly accompanied by a slight change in syntactic properties. And if we want to describe valency really explicitly, even such small shifts force us to introduce new LUs and to make the lexicon bigger than we would intuitively expect. Just two notorious examples for the beginning:

(34) The sun.ACT radiates heat.PAT
Heat.ACT radiates from the sun.DIR1.

(35) He loaded the truck.PAT with hay.MEANS.
He loaded the hay.PAT on the truck.DIR3.

Clearly, different frames (containing different functors) are instantiated in both pairs. Thus we have to have (at least) two different LUs for 'to heat' and (at least) two for 'to load' in the lexicon. But we cannot further ignore the fact that the LUs are obviously semantically related. According to [Levin, 1993], we will use the term alternation for naming such relation between two similar LUs.[31] However, unlike Levin, we don't want to study alternations because of building verb classes (although it might be an important side effect in the future), but we use alternations primarily for decreasing lexicon redundancy. The point here is that instead of having two unrelated LUs in the lexicon, it is more economic (less redundant) to store only one of them, but together with the information about applicability of the appropriate alternation on this LU. Thus the second LU does not have to be physically present in the lexicon, but can be generated 'on demand' by applying the alternation on the first LU.

Of course, this solution is more economic only for those alternations which are applicable on a reasonably big number of LUs (obviously it makes no sense to spend time on capturing alternations which occur only once or twice in the lexicon). Moreover, if we want to claim that no information was lost from the lexicon, then we can use only perfectly regular alternations.

Such alternation-based approach changes the conceptual view on the lexicon. Originally, the lexicon contained a mere list of LUs for each lexeme. Now, we group the LUs into *LU clusters*,[32]—each cluster contains a *basic LU*, which has to be physically stored in the lexicon, and possibly a number of

[31] Remark on terminology: Also other terms came into play here: the term *diathesis* was introduced in Meaning-Text Theory for a similar concept, but this term is mostly used only for alternations having to do with subject (especially various types of passivizations) in the present. [Daneš, 1985] used the term *hierarchization*, but this term did not get used by other researchers. That is why we adhere to the term alternation, although it might be misleading too: occasionally it is used in situations where there are more possible surface realizations of a surface slot filler (in order not to confuse the reader, we will not use it in this sense throughout this thesis).

[32] The term cluster was suggested to us by Petr Strossa.

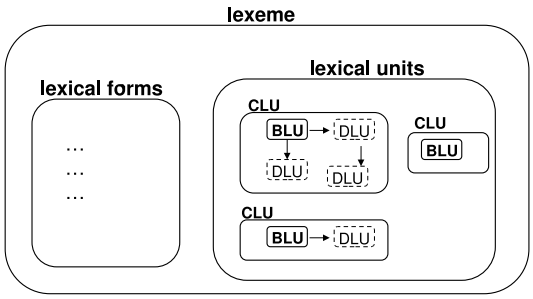


Figure 5.9: Alternation-based lexeme model (CLU stands for cluster of lexical units, BLU for basic lexical unit, and DLU for derived lexical unit.

derived LUs, which are present only virtually in the lexicon. The alternation-based model is depicted in Figure 5.9.

Thus the LU cluster can be understood as an oriented graph (with LUs as nodes and alternations as oriented edges) with one distinguished node from which there is an oriented path to all remaining nodes of the same cluster. The graph is not necessarily trivial: it does not have to have a simple 'star' topology, because some alternations can be composed together, and it does not even have to be a tree, because in some cases the composition of alternations might be commutative.

The difficult question arises which LU from the given cluster should be considered as the basic one, which was observed already in [Daneš, 1985], page 57. We do not have a satisfactory answer either. The formal criterion that all other LUs within the cluster should be reachable by applying the alternation(s) on the basic LU does not help much, since it is obvious that the choice of the basic LU is dependent on the inventory of alternations; if it would be possible to introduce for each alternation also its inverse, the choice would be completely arbitrary. But anyway, as soon as there is at least one (oriented) cycle in the cluster, the choice is arbitrary to some extent too. Therefore the basic LU cannot be unambiguously chosen without using some conventions.

5.11.2 Threefold effect of alternation

The relation of alternation can be seen as a unary function, which takes a LU as an argument and returns another LU as a result. The former is called *input LU* and the latter is called *output LU* in the following text. These two terms must be distinguished from the terms *basic LU* and *derived LU*, because (1) alternations can be composed, and thus a derived LUs becomes the input LU, (2) the graph of LU cluster might contain cycles (especially when considering inverse alternations), and thus the basic LU becomes output LU after applying an alternation on a derived LU.

In our approach, the effect of alternation is manifested by (at least one of) the following ways:[33]

- the output LU has a different valency frame from the input LU,
- the output LU has different constraints on the possible forms of frame evoker,
- the output LU has a different lexical meaning from the input LU.

As for the first two aspects, alternations can also be understood as lexically conditioned grammar rules, and thus should be conceptually situated on the border between grammar and lexicon.

Now may briefly illustrate the effects of various types of alternations on examples from Czech.[34]

(36) Jan miluje Marii.
John loves Mary.

(37) Marie je milována Janem.
Mary is loved by John.

In FGD, it was implicitly presumed that active and passive voice are instances of the same valency frame. However, in agreement with [Babby, 1998] we believe that active and passive voices should not be priviledged to other alternations (diatheses), and we treat them as two LUs related by the active/passive alternation.

(38) Naložil vůz senem.
He loaded the truck.PAT with hay.MEANS.

(39) Naložil seno na vůz.
He loaded the hay.PAT on the truck.DIR3.

The spray/load alternation (38) and (39) works in Czech in the same way as in English. The difference in the valency frame is obvious: the frame of the input LU contains a dependent in instrumental, whereas the frame of the output LU contains directional free modifier. The difference in the lexical meaning is (besides swapping the content of the second and third valency slot) the fact that the input LU (unlike the output one) imply that the truck was full of hay. As for the constraints laid on the frame evoker, there is no difference in this type of alternation.

[33] It should be reminded here that each alternation should be applicable on a whole class of LUs and that its manifestations must be regular (omitting these conditions could lead to an absurd claim that the whole language system can be represented by one single LU and all the rest are only alternations).

[34] Many more inspiring examples for Czech can be found also in [Daneš, 1985], pages 51-63.

(40) Vyšel na kopec.DIR3.
He climbed up to (the top of) the hill.

(41) Vyšel kopec.PAT.
He climbed the hill.

The difference between (40) and (41) is that the latter one expresses the climbing person conquered the hill as a whole (or at least a significant part of it), and not that he just stepped on the very top. Note that this alternation can be applied only on the LU with the sense 'to go up' and not on the other LU with the sense 'to go out':

(42) Vyšel na dvůr.DIR3.
He went out to the yard.

(43) *Vyšel dvůr.
*He went out the yard.

Thus it is obvious that the given alternation is limited only to one of the senses of polysemous prefix *vy-*.

(44) Oloupal kůru.PAT z pomeranče.DIR1.
He carved the skin off the orange.

(45) Oloupal pomeranč.PAT.
He carved the orange.

The difference caused by the alternation illustrated by (44) and (45) is that the free modifier in the input LU shifted to the patient position in the output LU, leaving no place for expressing what was carved from the orange surface.

(46) Ten zážitek učinil Jana.PAT dospělým mužem.EFF.
That experience made John an adult man.

(47) Ten zážitek učinil z Jana.ORIG dospělého muže.PAT.
That experience made an adult man from John.

In examples (46) and (47) the contents of the second and third valency slots are interchanged, and the surface form of the third slot is changed.

(48) Jan líbá Marii.
John kisses Mary.

(49) Jan se líbá s Marií.
John kisses with Mary.

(50) Jan a Marie se líbají.
John and Mary kiss (each other).

(51) Dvojice se líbala.
The pair was kissing.

Examples (48)–(51) illustrate reciprocity. The kissed person can either be expressed in accusative, or in the prepositional group with preposition *s*, or – if she is in the symmetric relation to the other person – the two can be coordinated (and then the frame evoker contains the reflexive particle), or the both participants can be 'hidden' inside a noun with the meaning of a pair or group. Of course, the (50) and (51) are instances of the same LU.

(52) Marie si plete Jana s Petrem.
Mary confuses John with Peter.

(53) Marie si plete Jana a Petra.
Mary confuses John and Peter.

(54) Jan nerozliší třešně od višní.
John does not recognize sweet cherries from sour cherries.

(55) Jan nerozliší třešně a višně.
John does not recognize sweet cherries and sour cherries.

The alternation shown in examples (52)–(55) is in a sense similar to reciprocity, but the second and third slot are merged into one (instead of the first and second slots), and since the subject is left untouched by this alternation, there is no added reflexive particle in the output LU frame evoker.

(56) Marie spí.
Mary sleeps.

(57) Marii se tu spí dobře.
lit. Mary.dat refl. here sleeps.sg.neut well.

Examples (56) and (57) illustrate what is called dispositional modality in FGD. Actor is expressed by dative in the output LU (whereas it was nominative in the input LU), evaluative adverbial is obligatory in the valency frame, the reflexive particle *se* was added to the frame evoker, and the meaning of the output LU contains a modality feature (the actor is able to do the activity to the degree expressed by the adverbial).

Dispositional modality is not applicable to all verbs in Czech, e.g. *cítit* (to feel).

(58) Marie malovala obrázky na stěnu.
Mary drew pictures on the wall.

(59) Marie pomalovala stěnu obrázky.
Mary drew the wall with pictures.

Examples (58) and (59) show that even relation between LUs from different lexemes can be treated as alternations. In this case, what was expressed as free modifier in (58) is patient in (59), and what was patient in (58).

(60) Marie cestuje do Prahy.
Mary travels to Prague.

(61) Marie se do Prahy nacestuje hodně.
Mary will often travel to Prague.

Examples (60) and (61) can be treated as manifestations of another inter-lexeme alternation. The extremely productive prefix *na-* combined with reflexive particle *se* expresses that the activity takes a long time or is often repeated. A new slot corresponding to an expected quantification or frequentative adverbial is added to the frame.

5.11.3 Minimal and expanded form of the lexicon

According to the alternation-based model, the lexicon (in its minimal form) contains only basic LUs with associated lists of applicable alternations. However, there are various situations in which it could be useful to physically add the derived LUs into the lexicon too:

- if we want to attach some non-predictable information to the derived LUs, e.g. number of occurences of a given derived LU in a corpus (especially if gained from human annotation), translational equivalents in other languages, synonyms etc.,
- if we want to perform automatic frame disambiguation (deciding which frame was used in the given sentence) on a bigger piece of data, then we need to have an access to all LUs, not only to the basic ones. Obviously it would be highly inefficient to generate the derived LUs again and again for each occurence of a given lexeme; it would be much faster to generate each of them only once at the beginning and to store them in the lexicon.

The consequences for the lexicon design are the following:

- minimal and expanded forms of the lexicon should be distinguished; if no new LU can be added by applying the alternations, then the lexicon is totally expanded; if it contains only a subset of possible derived LUs, then it is partially expanded,
- an automatic procedure should be implemented for converting the lexicon from its minimal form into its totally expanded form,
- the data representation of the basic and derived LUs should be as similar as possible, so that they can be accessed in a similar way (however,

the content of the derived LUs can be hardly fully equivalent to that of basic LUs, because it is extremely difficult to automatically generate example usages of the derived LUs).

So far, the alternations are not annotated in VALLEX; to our knowledge there is no in-depth study of alternations for Czech comparable to that in [Levin, 1993] for English, therefore it was not possible to start a large-scale manual annotation of alternations in VALLEX. However, we believe that the alternation-based model sketched in this section offers an efficient and unified mechanism for capturing many diverse types of regularities in the language and thus will sooner or later lead to a significant reduction of the lexicon redundancy. And hopefully also to a better insight into lexical semantics of verbs.

Chapter 6

Annotation Scheme of VALLEX

Talkie Toaster: Given that God is infinite,
and that the Universe is also infinite,
would you like a toasted tea cake?
[Red Dwarf]

In this chapter we look at the valency lexicon VALLEX from the practical point of view. All main software components of the dictionary production system will be briefly described, as well as the data formats and annotation processes.

6.1 Editing environment and primary annotation format

Software environment for manual annotation of lexicon entries is one of the most important parts of the dictionary production system, since its properties determine the speed and effectivity of annotation, and can influence also the number of annotation errors. That is why we paid a special attention to choosing the optimal solution.

First, we have developed a conventional relational MS Access database with a special graphical user interface. However, it was clear quite soon that such a solution is too cumbersome: the navigation through the lexicon was not comfortable (the mouse had to be used often for pressing buttons, closing windows, scrolling lists etc.), and any subtle change in the annotation format (leaving aside adding new features) required intervention into the database source codes and redistribution of the compiled file.

Another solution was based on the idea of editing directly the XML representation of the lexicon in a text editor. This was more comfortable for the annotators, because it allowed for fast navigation and simple search through the file, cut-and-past annotation of frequent patterns etc. Moreover, the annotated XML file could be immediately and easily transformed into HTML via XSL transformation, and viewed in a web browser. However, from the visual point of view, the annotated entries were too long because of lots of XML tags, which made them difficult to read.

The third approach, which we use up to now and find it optimal for our needs, is based on a combination of a special line-oriented plain-text data

format with the text editor WinEdt.[1] Due to the simple notation convention in the data format, it was possible to create a syntax-highlighting mode for WinEdt, which visually distinguishes different parts of the entry. Thus the navigation through the lexicon and the manipulation with the entries is quite comfortable. Moreover, some types of annotation errors are noticed by the annotator instantly and without any effort, e.g. because of unusual combination of colors.[2]

The lexicon data are not merged in one large text file, but are divided into several smaller files according to the (dominating) semantic features of the individual verbs. Such division enables easy workload distribution among the annotators.

Currently, the annotated text files are uploaded by the annotators into a CVS directory. Of course, it has to be ensured that at most one person can change a given file at a time.[3]

In the following list, we give a simplified description of the notation rules used in the text format:

- anything between the "#" symbol and the end of line is a comment and is not further processed,
- any sequence of spaces is treated as a single space; a space at the line beginning is ignored,
- each lexeme entry contains a line starting with "*" followed by a list of lemmas representing the given lexeme, and a sequence of lexical unit (LU) entries,
- each LU entry consists of a line with lemmas, a line with frame, and a sequence of lines with frame attributes,
- the line with lemmas starts with the "~" symbol followed by a list of lemmas, each of them preceded with a shortcut representing its morphological aspect ("dok:" for perfective, "nedok:" for imperfective, "nás:" for iterative, "dokned:" for biaspectual),
- the line with frame starts with the "+" symbol followed by a sequence of frame slots,
- each frame slot matches the pattern "FUNC(surf-real;type)", where FUNC is a functor, surf-real is a list of surface realizations separated by comma (the list can be empty, if the surface realizations are implicitly implied), and type distinguishes between obligatory, optional and typical slots,

[1] http://www.winedt.com

[2] It should be noted that our approach is not limited to one particular editor. As it was shown in [Hlaváčková and Horák, 2005], our solution can be easily transfered to the editor VIM.

[3] It would not be difficult to implement a distributed system allowing synchronous work on the same data with preventing the possible collisions in the same time, but it would not make the situation much easier because most annotators cannot work on-line.

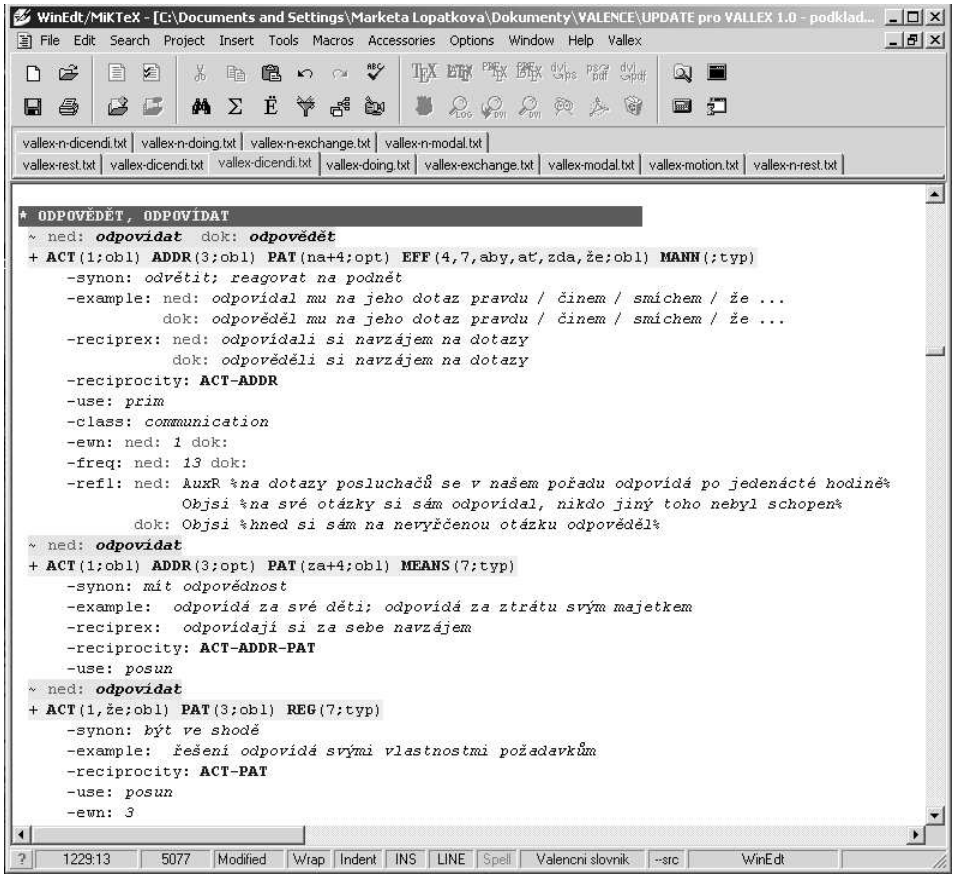


Figure 6.1: Highlighting mode in Winedt text editor as the environment for writing lexicon entries.

- the line with frame attribute starts with the pattern "-attrname:" (where attrname is the name of the attribute) followed by the value of the attribute; when necessary, the value is separated into parts related to individual aspectual counterparts (e.g. examples of usage contain the verb form and thus it has to be different for the two counterparts); the set of attribute names is left open, and thus new or temporary attributes can be easily added to to entries without reimplementing anything,
- by default, the frame attribute value continues on the following line, unless it starts with one of the special symbols ("-","+","*","~")

This format is easily parsable and convertible into other formats. We use a simple finite state automaton implemented in Perl for converting the data into XML.

6.2 Selection of the lexical stock

There are around 19000 Czech verbs in the Czech National Corpus (CNC), and maybe twice as many in the whole contemporary Czech.[4] Such amount is not directly manageable by a small lexicographic team, therefore some selection criteria have to be determined at the very beginning.

We naturally select corpus frequency as the main criterion: verbs with high frequency should get processed first. In this way, the NLP usability in terms of corpus coverage grows fastest.[5] We used the frequency lists gained directly from the Czech National Corpus,[6] because no electronic frequency lexicon was available when we started building VALLEX. Nowadays, the frequency dictionary [Blatná et al., 2004] could be also used.

As we mentioned in 4.1, it might be confusing to use directly the term verb. When speaking about corpus occurrences, we use here the term *m-lemma*, which denotes a single morphological lemma (as contained in the corpus), regardless of whether it is a reflexive or not (which is highly non-trivial to detect in a syntactically unannotated corpus). The cumulative coverage of Czech verb m-lemmas is given in Figure 6.2.

In the pilot version of VALLEX, we started with around 100 m-lemmas (leading to around 180 verb entries, because of the distinguishing of reflexives). Later, several new sets of m-lemmas (several hundreds of m-lemmas

[4] This is especially due to productive word-formative language means such as prefixing.

[5] Although this is a widely preferred approach in contemporary lexicography, one should keep in mind that the created data resource is skewed (not representative), for instance due to the fact that the most frequent verbs tend to have higher polysemy in comparison with the less frequent ones, and thus the average polysemy in the created lexicon is higher than in the virtually complete lexicon.

[6] Due to the tagging errors in CNC, verbs such as "pět" (to sing, archaic) as a wrongly tagged numeral "pět" (five), or "telit" (to calve) as a wrongly tagged abbreviation "tel." (telephone) appear among the most frequent verbs in the frequency list.

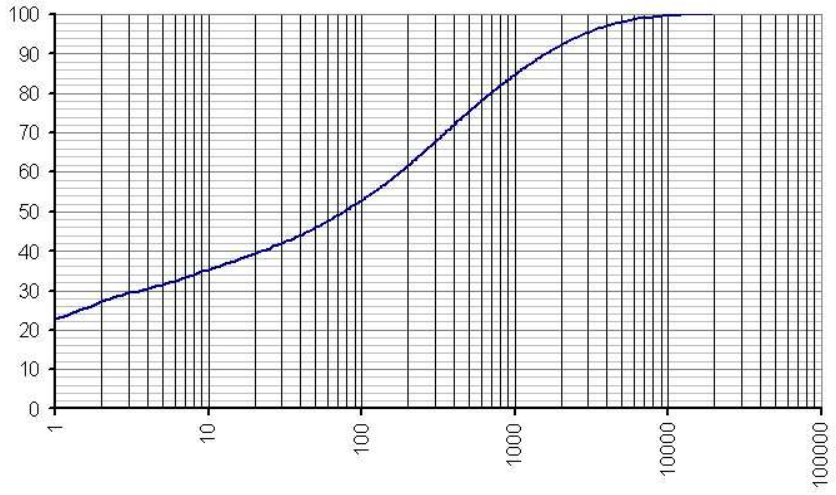


Figure 6.2: Corpus coverage of Czech verb m-lemmas (based on Czech National Corpus, subcorpus SYN2000). Logarithmically scaled horizontal axis corresponds to the sequence of verb m-lemmas sorted according to their corpus frequency, whereas the vertical axis corresponds to the cumulative number of occurrences divided by number of all occurrences of verbs.

each) were gradually added. The first public release, VALLEX 1.0, contained around 1000 m-lemmas (1400 "verbs"). In the present version of VALLEX there are around 1800 m-lemmas (more detailed quantitative information is given given in Section 6.9).

It is important to note that frequency is not the only criterion for adding a verb into VALLEX: once an m-lemma is added, the m-lemma of the aspectual counterpart is added too, no matter how frequent it is.

6.3 WWW interface for searching the text format

In simple cases, the text files with manually annotated lexicon entries can be searched directly in the text editor in which they are created (especially if the editor supports searching regular expressions), but this is not possible in more complex queries, e.g. those which specify conditions for different parts of entries. That is why we have developed a simple search engine for querying the lexicon text files.

Our solution was implemented using the CGI (Common Gateway Interface) technology. The user can easily specify the query in a HTML form using his/her WWW browser. When the user submits the query, the web server (Apache Web Server in our case)[7] executes a CGI script which we

[7] http://www.apache.org

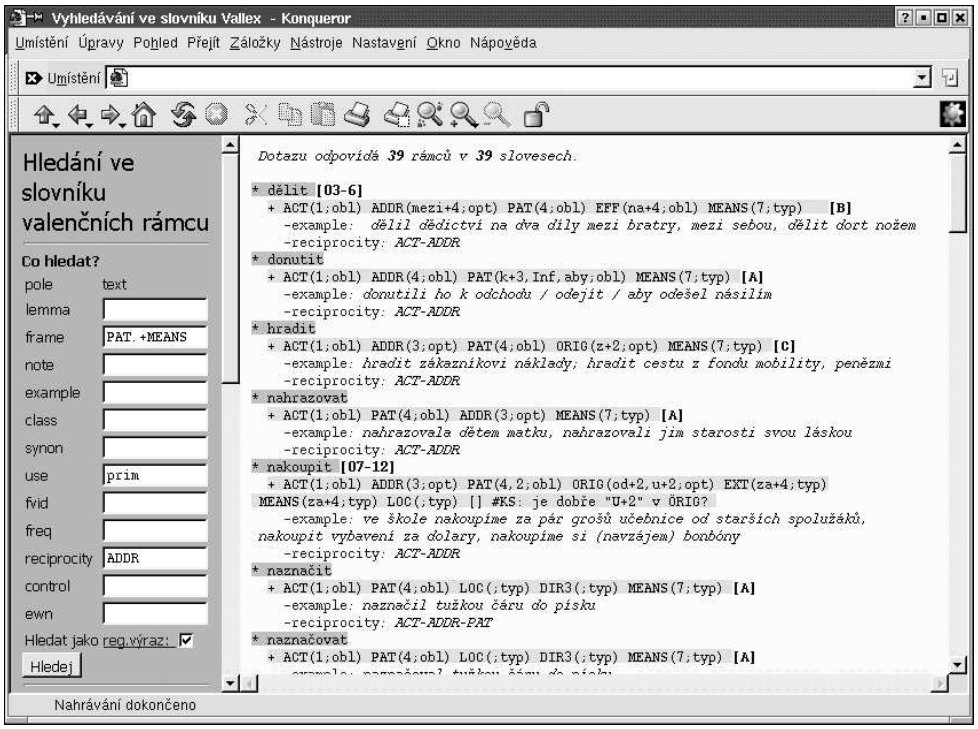


Figure 6.3: WWW interface for searching the lexicon text files.

have implemented in Perl; the script loads (and parses) the lexicon text files from the file system, evaluates the conditions specified in the query and returns the result to the web server. It sends it back to the user's browser.

A screenshot of the WWW interface is characterized in Figure 6.3. The present version of the interface allows to specify the following conditions:

- a regular expression can be specified for each part of the lexicon entry separately; if regular expressions are filled in more fields, only the entries matching all the expressions are returned,
- one can select what part of the entry should be displayed in the resulting response,
- one can restrict the set of files which are to be searched,
- instead of obtaining the individual entries found, the user may obtain also selected basic distributional properties (the frequency list of functors, the frequency list of surface forms, or their combinations etc.).[8]

[8]This is especially useful when checking the consistency of the lexicon, applying the common-sense rule "the less frequent the more suspicious."

6.4 Annotation process

In this section we briefly summarize how new entries are added to VALLEX.

First, a new portion of so far unannotated verb m-lemmas (usually in the size of five hundred items) is picked upc according to their CNC frequency. Note that these new m-lemmas do not have to form a continuous segment within the frequency list, since some of the less frequent m-lemmas are already present in VALLEX because of their more frequent aspectual counterparts. The list of the new m-lemmas is stored into a text file which constitutes the nucleus of the new lexicon files.

Second, aspectual counterparts are added to m-lemmas.

Third, the text file with the aspectual pairs is divided into several text files, so that the new files contain semantically related verbs (in a very loose sense). This seems to be a more practical organization than e.g. cutting the alphabetical list.

Fourth, a set of around 100 sentences is automatically extracted from CNC for each verb and stored into HTML format, so that the annotators can instantly observe the verbs' behavior in authentic language material, without querying the corpus themselves, again and again.

Fifth, the files are distributed among annotators and they start creating the entries, verb by verb (reflexives must be added at this step!), sense by sense.[9] Although they physically create the entries from scratch,[10] they are heavily using other language resources: besides the already mentioned CNC, they also study the entries in other dictionaries, especially [Svozilová et al., 1997] and [SSJČ, 1978].

When finishing the first version of the new files, the new entries are intensively tested in order to detect (and correct) annotation errors and inconsistencies. The tests can be tentatively classified into two classes:

- technically-oriented tests – it is necessary to check whether the files fulfil the notation convention (this is best tested when converting the data into XML), whether there are any spelling errors, whether no verb occurs twice in the lexicon, etc. These tests can be performed fully automatically

- linguistically-oriented tests – it should be tested whether the entries are complete, especially whether all verb senses are captured, whether all slots and their frames are present, etc. Moreover, it should be checked whether similar verbs are annotated in an analogical way. These tests cannot be performed automatically and require very huge concentration. In some situations, it is better to completely rewrite

[9] We put emphasis on processing each verb in all its senses in VALLEX.

[10] At the beginning, we intensively experimented with automatically pre-generated entries (based e.g. on lexicon BRIEF), but to our experience it made the annotation process slower and more erroneous in comparison to writing the entries from scratch.

the whole entry. Using more corpus examples might be useful in such cases, especially in the case of highly polysemous verbs.

This last step is a virtually never ending process[11] and thus the VALLEX lexicon is subject to a continuous change. Of course, the end users of VALLEX can be hardly supposed to accept this view. However, we know that no final perfect version can ever exist, but it is possible to release the data in a discrete sequence or "frozen" versions of the data. So far, we have created one publicly available release of VALLEX described in the next section.

6.5 Release and distribution of VALLEX 1.0

The first publicly available full-fledged release of our valency lexicon is called VALLEX 1.0 and was issued in autumn 2003. It contained around 1400 Czech verbs with around 4000 valency frames. Besides the data, VALLEX 1.0 contains also a detailed documentation of the lexicon, including selected publications.

In order to satisfy different needs of different potential users, we distribute the lexicon in the following three formats:

- *XML version.* The data in the primary text format[12] was converted into a single XML file, the structure of which is contained in Appendix B. This format was used for converting the lexicon into the remaining two formats, as it is given in Figure 6.4. Of course, the XML format is intended especially for programmers, whereas the following two formats do not require any special computer skills.

- *Browsable version.* The HTML version of the data allows for an easy and fast navigation through the lexicon in usual WWW browsers (see the screenshot in Figure 6.5). Verbs and frames are organized in several ways, the selectional criterion can be chosen by clicking the button in the topmost frame in the browser. For instance, the user can easily list the frame containing a certain functor or a certain surface realization, view all frames belonging to a certain class, or view all perfective or imperfective verbs, and many other. This browsable version consists of more than 2000 pre-generated static HTML pages. The graphical layout is rendered via CSS (Cascade Style Sheets) technology.

- *Printable version.* VALLEX 1.0 entries were also converted into a format feasible for printing. The sample from the printed version of the

[11] Not only because of correcting the errors, but also due to application of new achievements in the background theory.

[12] The files in the primary text format were not distributed within VALLEX 1.0, since they are intended to be used for development purposes, but are not supposed to be used directly by the end user.

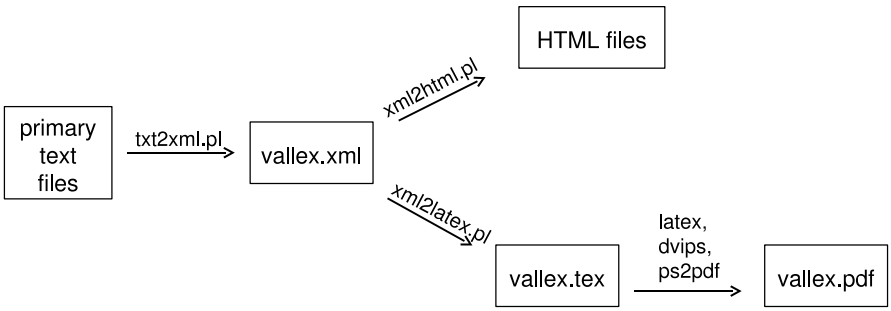


Figure 6.4: Conversion of the VALLEX 1.0 formats for the release purposes.

lexicon is given in Figure 6.6. We tried to keep a graphical layout visually similar to that in HTML, but make it more compact to save space (yet the printed version is 200 pages long). We used the document formatting system LaTeXfor creating the printed version.

All conversion tools (from the primary format into XML, from XML to HTML, from XML to LaTeXsource code) were implemented in Perl.

As for the distribution of VALLEX 1.0, we have decided to make it publicly available[13] on the Internet. The main site of VALLEX 1.0 is

http://ckl.mff.cuni.cz/zabokrtsky/vallex/1.0/

After registration on this site, the user obtains an email specifying where he/she can download the compressed dictionary from. Usage of the lexicon is limited to non-commercial purposes, as it is stated in the license agreement.[14].

At the time of finishing this thesis (summer 2005), there are more than one hundred registered users of VALLEX 1.0.

Besides the electronic distribution, VALLEX 1.0 (with all entries in the printable format) was issued as a technical report [Lopatková et al., 2003].

Since the release of the first version, the lexicon was further developed both in qualitative and quantitative aspects. A new XML representation described in the next section has been designed, and many previously unseen lexemes have been annotated. The new internal version of the lexicon is denoted as VALLEX 1.5 and its quantitative properties are presented in Section 6.9.

[13] In this aspect, we share the view of [Koster and Gradmann, 2004]: "We defend the thesis that basic linguistic resources, especially when developed largely with public money, should be made freely and openly available to the public."

[14] http://ckl.mff.cuni.cz/zabokrtsky/vallex/1.0/doc/license.html

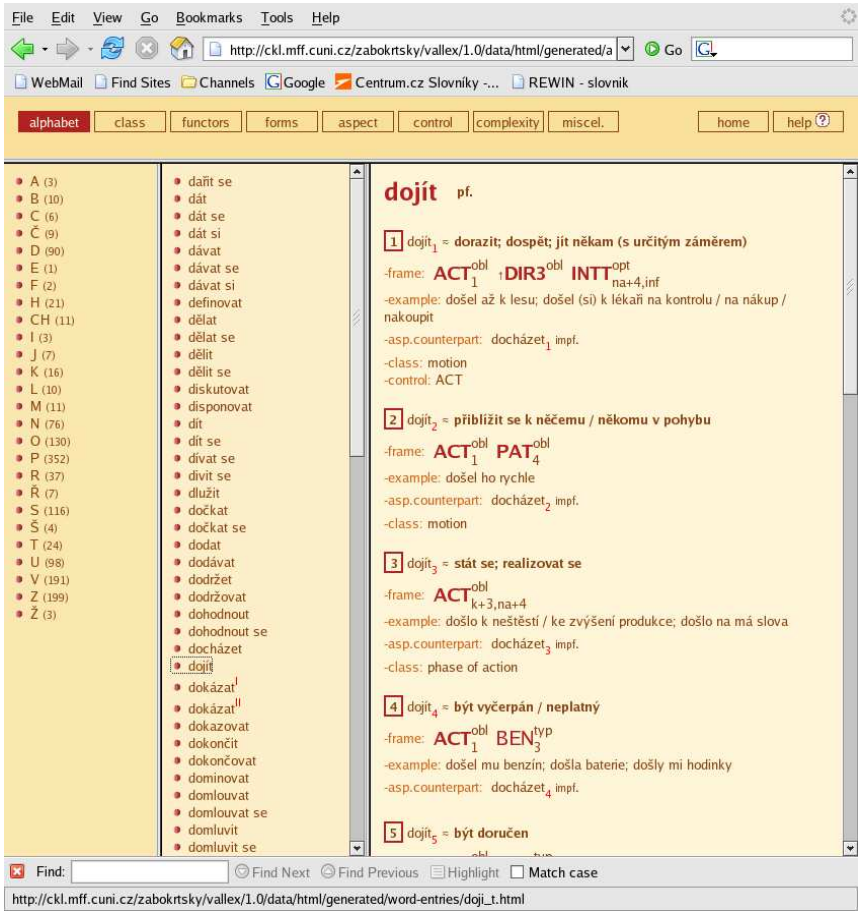


Figure 6.5: Sample from the browsable version of VALLEX 1.0.

2 hrát si$_2$ ≈ **předstírat** (idiom)
–frame: $\mathbf{ACT}_1^{obl}$ $\mathbf{PAT}_{na+4}^{obl}$
–example: *Petr si hraje na machra*
–asp.counterparts: hrávat si iter.

hrozit impf.

1 hrozit$_1$ ≈ **vyhrožovat**
–frame: $\mathbf{ACT}_1^{obl}$ $\mathbf{ADDR}_3^{obl}$ $\mathbf{PAT}_{7,že}^{obl}$
–example: *hrozil nám udáním / že nás udá*
–asp.counterparts: hrozívat iter.
–class: communication

2 hrozit$_2$ ≈ **vyhrožovat gestem**
–frame: $\mathbf{ACT}_1^{obl}$ $\mathbf{PAT}_3^{opt}$ $\mathbf{MEANS}_7^{typ}$
–example: *hrozil nám rukou*
–asp.counterparts: hrozívat iter.

3 hrozit$_3$ ≈ **blížit se**
–frame: $\mathbf{ACT}_2^{obl}$ $\mathbf{PAT}_{1,že}^{obl}$ $\mathbf{LOC}^{typ}$
–example: *hrozil mu neúspěch; v Mongolsku hrozí hladomor*
–asp.counterparts: hrozívat iter.

hrozit se impf.

1 hrozit se$_1$ ≈ **obávat se; děsit se**
–frame: $\mathbf{ACT}_1^{obl}$ $\mathbf{PAT}_{2,aby,že}^{obl}$
–example: *hrozil se neúspěchu*
–asp.counterparts: hrozívat se iter.

hýbat impf.

1 hýbat$_1$ ≈ **pohybovat; měnit polohu něčeho**
–frame: $\mathbf{ACT}_1^{obl}$ $\mathbf{PAT}_{7,s+7}^{obl}$
–example: *hýbat kliskou / rukou / s nábytkem*
–asp.counterparts: hýbnout$_{1\ pf.}$

2 hýbat$_2$ ≈ **vzbuzovat zájem / rozruch** (idiom)
–frame: $\mathbf{ACT}_1^{obl}$ $\mathbf{PAT}_7^{obl}$
–example: *nové myšlenky hýbou světem*

hýbat se impf.

1 hýbat se$_1$ ≈ **pohybovat se; měnit polohu**
–frame: $\mathbf{ACT}_1^{obl}$ $\mathbf{LOC}^{typ}$ $\uparrow\mathbf{DIR}^{typ}$
–example: *Nehýbejte se!; větev se hýbá ve větru*
–class: motion

hýbnout pf.

1 hýbnout$_1$ ≈ **pohnout; změnit polohu něčeho**
–frame: $\mathbf{ACT}_1^{obl}$ $\mathbf{PAT}_{7,s+7}^{obl}$
–example: *hýbnout hlavou / se skříní*
–asp.counterparts: hýbat$_{1\ impf.}$

CH

charakterizovat biasp.

1 charakterizovat$_1$ ≈ **popsat, popisovat; vystihnout, vystihovat**
–frame: $\mathbf{ACT}_1^{obl}$ $\mathbf{PAT}_4^{obl}$ $\mathbf{MEANS}_7^{typ}$ $\mathbf{COMPL}_{jako+4}^{typ}$
–example: *problém charakterizoval těmito slovy; ta vlastnost ho dost charakterizuje; charakterizoval přítele jako dobráka*
–class: communication

chodit impf.

1 chodit$_1$ ≈ **pohybovat se pomocí nohou; přemísťovat se (s nějakým záměrem)**
–frame: $\mathbf{ACT}_1^{obl}$ $\mathbf{INTT}_{na+4,inf}^{opt}$ $\mathbf{MANN}^{typ}$ $\uparrow\mathbf{DIR}^{typ}$
–example: *chodit domů pěšky; chodit od hospody k hospodě; chodit rychle; dítě už chodí; chodí stejně (ale jako Jirka.CPR); chodit na borůvky / na nákup / nakupovat; chodit k lékaři na kontroly*
–asp.counterparts: chodívat iter.
–class: motion
–control: ACT

2 chodit$_2$ ≈ **absolvovat chůzí**
–frame: $\mathbf{ACT}_1^{obl}$ $\mathbf{PAT}_4^{obl}$
–example: *chodit pochod*
–asp.counterparts: chodívat iter.
–class: motion

3 chodit$_3$ ≈ **být doručován** (idiom)
–frame: $\mathbf{ACT}_1^{obl}$ $\mathbf{BEN}_{3,pro+4}^{typ}$
–example: *pošta chodí i v neděli;'chodí špatné zprávy z Rwandy*
–asp.counterparts: chodívat iter.

4 chodit$_4$ ≈ **fungovat** (idiom)
–frame: $\mathbf{ACT}_1^{obl}$ $\mathbf{MANN}^{typ}$
–example: *chodit bez chyby o stroji; ten stroj už chodí*
–asp.counterparts: chodívat iter.

5 chodit$_5$ ≈ **ujídat** (idiom)
–frame: $\mathbf{ACT}_1^{obl}$ $\mathbf{PAT}_{na+4}^{obl}$ $\uparrow\mathbf{DIR}^{typ}$
–example: *chodit na hrušky / na cukroví do komory*
–asp.counterparts: chodívat iter.

6 chodit$_6$ ≈ **být upraven** (idiom)
–frame: $\mathbf{ACT}_1^{obl}$ $\mathbf{PAT}_{adj-1}^{obl}$
–example: *chodit otrhaný; chodí na bál přestrojená*
–asp.counterparts: chodívat iter.

7 chodit$_7$ ≈ **mít partnera** (idiom)
–frame: $\mathbf{ACT}_1^{obl}$ $\mathbf{PAT}_{s+7}^{obl}$
–example: *chodit s někým*
–asp.counterparts: chodívat iter.
–class: social interaction

8 chodit$_8$ ≈ **být oblečen** (idiom)
–frame: $\mathbf{ACT}_1^{obl}$ $\mathbf{COMPL}_{jako+1,za+4}^{typ}$
–example: *chodí jako maškara / za maškaru o masopustu*
–asp.counterparts: chodívat iter.

Figure 6.6: Sample from the printable version of VALLEX 1.0.

6.6 VALLEX XML, version B

The XML data format used in VALLEX 1.0 and documented in Appendix B was not intended to be the ultimate and the only XML representation of VALLEX; it was rather a pilot study. In this section a newer version is presented, which is in better agreement with what was said about the lexicon in Chapter 4 and 5. In order to differentiate the numbering the VALLEX data versions from the numbering of the versions of the VALLEX XML formats, we denote this new format with the letter 'B'.

Besides more or less marginal changes (such as adding identifiers, renaming elements, adding intermediate elements etc.), there are three very important differences between these two formats:

- Reflexive particles are strictly separated from m-lemmas in the B-version, and are associated only with the given lexeme (i.e., with the set of its lexical forms) as a whole, not with its individual m-lemmas.
- Aspectual counterparts are merged into single lexemes in the B-version. This requires adding two new mechanisms: one for co-indexing examples and glosses with the relevant m-lemmas, and one for capturing the situation when a certain LU is associated only with one of two aspectual counterparts.
- The B-version format is ready for the alternation-based lexicon model (although in praxis in is has not been filled with alternations and derived LUs yet).

In the rest of this section we present commented fragments from the Document Type Definition of the B-version of VALLEX XML.

In the B-version of the VALLEX XML format, the root element *vallex_b* consists of two parts – `head` contains the general information about the lexicon, whereas `body` contains the data:

```
<!ELEMENT vallex_b (head, body)>
```

In the `head` part, the title of the presented lexical resource is specified (VALLEX - Valency Lexicon of Czech Verbs), the version of data and the date when this XML file was generated (in the following sections we work with the version 1.5 created in June 2005), names of the authors (Markéta Lopatková, Zdeněk Žabokrtský, Karolína Skwarska, Václava Benešová), and a short description of VALLEX:

```
<!ELEMENT head (title, version, last_change, authors, description)>
<!ELEMENT title (#PCDATA)>
<!ELEMENT version (#PCDATA)>
<!ELEMENT last_change (#PCDATA)>
<!ELEMENT authors (author+)>
<!ELEMENT author (#PCDATA)>
<!ELEMENT description (#PCDATA)>
```

The `body` part of the lexicon consists of two elements:

```
<!ELEMENT body (entries, relations)>
```

The element `entries`, which contains lexemes and their LUs, whereas the element `relations` is intended to contain the links between lexemes or between lexical units, especially for representation of word-formative relations (e.g. due to prefixing) and lexical functions (as outlined in Sections 4.7 and 4.9). The development of the representation of such relations has not been stabilized so far, therefore the element `relations` is left empty in the presented DTD:

```
<!ELEMENT relations EMPTY)>
```

The lexeme entries are not inserted directly into the element `entry`, but are first clustered within the element `lexeme_cluster`:

```
<!ELEMENT entries (lexeme_cluster+)>
<!ELEMENT lexeme_cluster (lexeme+)>
```

so that the reflexive and irreflexive lexemes (see Section 4.2) sharing the same m-lemma(s) (such as three lexemes *brát/vzít*, *brát si/vzít si*, *brát se/vzít se* sharing m-lemmas *brát* and *vzít*) get closer in the data structure, as they are supposed to be processed together in many situations. For instance, in word sense disambiguation task, once a given m-lemma occurs in a sentence, both reflexive and irreflexive lexemes come into play, no matter whether there is any reflexive particle in the outer shape of the sentence or not.

Now, the lexeme as defined in this thesis (Section 4.1) associates lexical forms with lexical units:

```
<!ELEMENT lexeme (lexical_forms, lexical_units)>
```

Besides that, each lexeme is equipped with the attribute `pos` specifying its part of speech (although there are only verbs in VALLEX at the current stage, the presented DTD is prepared also for nouns, adjectives and adverbs) and a unique identifier `id`.

```
<!ATTLIST lexeme
        pos (v|n|adj|adv) #REQUIRED
        id ID #REQUIRED
      >
```

A set of lexical forms manifesting the given lexeme is determined using three types of information: first, the m-lemmas are listed, second, the reflexive particle is specified (only in the case of reflexive lexemes), and third, other additional constraints can be specified:

```
<!ELEMENT lexical_forms
    ((mlemma|mlemma_variants)+, particle?, constraint*)>
```

The content of the element `mlemma` is the m-lemma itself (i.e. a sequence of letters in the national alphabet). For each m-lemma, its morphological aspect is obligatorily stored in the attribute `aspect`; in addition, in the case of homography (see Section 4.4) an Arabic number is stored into the attribute `homograph` to distinguish the given m-lemma from the other homograph:

```
<!ELEMENT mlemma (#PCDATA)>
<!ATTLIST mlemma
        aspect (pf|impf|biasp) #REQUIRED
        homograph CDATA #IMPLIED
        coindex CDATA #IMPLIED
      >
```

The attribute `coindex` has to be used because the aspectual counterparts are stored in the same lexeme (see Section 4.5), but certain items in the entries of the LUs within the lexeme must contain different values for these two counterparts. For instance, example sentences for *vzít si* and *brát si* (e.g. in the sense 'to marry') are simply different strings as they contain different inflected forms of different m-lemmas. Our solution is that we split the example into parts, which are co-indexed with the individual m-lemmas via the value of the attribute `coindex`. Glosses can be split in the same way as examples.

In the case of m-lemma variants (see Section 4.3), two or more m-lemmas are merged together into the `mlemma_variants` element:

```
<!ELEMENT mlemma_variants (mlemma+)>
<!ATTLIST mlemma_variant
        coindex CDATA #IMPLIED
      >
```

The element `particle` is used only within reflexive lexemes and contains the reflexive particle itself (simply said, either *si* or *se*):

```
<!ELEMENT particle (#PCDATA)>
```

The element `constraint` is supposed to be used in the case of other constraints on the surface form of the frame evoker (such as tense or negation, as discussed in Section 5.5). So far, such constraints have not been annotated in the VALLEX data, therefore we leave their representation partly underspecified at this moment:

```
<!ELEMENT constraint EMPTY>
<!ATTLIST constraint
        feature CDATA #REQUIRED
        value CDATA #IMPLIED
      >
```

Now we turn from the representation of lexical form to the representation of lexical units. Again, entries representing individual lexical units are not contained directly in the element `lexical_units`, but are clustered first (recall the alternation-based lexicon model presented in Section 5.11 and the distinction between basic LUs and derived LUs):

```
<!ELEMENT lexical_units (lu_cluster+)>
<!ELEMENT lu_cluster ((blu,dlu*)|dlu+)>
```

The element `blu` representing basic lexical units is structured as follows:

```
<!ELEMENT blu (lexical_forms?, frame, gloss, example,
               control?, class?, alternations?)>
```

and is of course associated with a unique identifier:

```
<!ATTLIST blu
    id ID #REQUIRED
>
```

If the set of lexical forms of the given LU is not identical with the set of lexical forms specified for the whole lexeme, then it can be re-specified: for the given LU, the content of the element `lexical_forms` embedded directly in the element `blu` overrides what was specified in `lexical_forms` embedded in *lexeme*. This is to be used especially when the given LU can be used only with one verb from the aspectual pair.

The most important part of each LU entry is of course its valency frame stored in the element `frame`. The valency frame consists of a (possibly empty) sequence of slots. In each slot, we specify properties related to both surface and deep valency frames, namely the list of possible surface forms (discussed in Section 5.6) represented by the elements `form`, the functor (discussed in Section 5.7) stored in the attribute `functor`, and the type of the slot, in which the dichotomy obligatory and optional (corresponding to the dialog test mentioned in 2.1.2 and enriched with the third value for free modifiers which are still related to the given LU in some specific way) is captured:[15]

```
<!ELEMENT slot (form*)>
<!ATTLIST slot
   functor (ACT|PAT|ADDR|EFF|ORIG|ACMP|ADVS|AIM|APP|APPS|ATT|BEN|CAUS|
         CNCS|COMPL|COND|CONJ|CONFR|CPR|CRIT|CSQ|CTERF|DENOM|DES|DIFF|
         DIR1|DIR2|DIR3|DISJ|DPHR|ETHD|EXT|FPHR|GRAD|HER|ID|INTF|INTT|
```

[15] We are aware of the fact that classification of frame slots with respect to their obligatoriness deserves much more attention than it gets in this thesis. First, the deep obligatoriness should be strictly distinguished from the surface obligatoriness (to our knowledge, the latter one has not been studied within FGD so far). Second, it is very likely that finer scales are necessary (e.g. [Somers, 1986] suggests six degrees of valency binding).

```
        LOC|MANN|MAT|MEANS|MOD|NA|NORM|PAR|PARTL|PN|PREC|PRED|REAS|
        REG|RESL|RESTR|RHEM|RSTR|SUBS|TFHL|TFRWH|THL|THO|TOWH|
        TPAR|TSIN|TTILL|TWHEN|VOC|VOCAT|SENT|
        DIR|OBST|RCMP)   #REQUIRED
  type (obl|opt|typ) #REQUIRED
    >
```

Surface forms captured by the element `form` are first classified by the attribute `type` (see Section 5.6). Its value implies which further attributes must be specified (for instance, the type `prepos_case` requires `case` and `prepos_lemma` to be filled):

```
<!ELEMENT form EMPTY>
<!ATTLIST form
  type (direct_case|prepos_case|subord_conj|rel|
        infinitive|adjective|phraseme_part) #REQUIRED
  case (1|2|3|4|5|6|7) #IMPLIED
  prepos_lemma (bez|do|jako|k|kolem|kvůli|mezi|místo|na|nad|na_úkor|
       o|od|ohledně|okolo|oproti|po|pod|podle|pro|proti|před|přes|
       při|s|u|v|ve_prospěch|vůči|v_zájmu|z|za|než) #IMPLIED
  subord_conj_lemma (že|aby|zda|ať|jak|až) #IMPLIED
  phraseme_part CDATA #IMPLIED
  to_be (0|1) "0"
>
```

Besides the valency frame, the element `blu` is obligatorily equipped with the elements `gloss` and `example`, and optionally with the elements `control`, `class`, `alternations`.

The element `gloss` contains a synonymous expression (or a close paraphrase) for the given LU. Its function is purely distinctive: it should help the user to quickly grasp which LU in the lexicon corresponds to which sense. As it was already mentioned, aspectual counterparts merged in one lexeme may require different glosses: in such a case the element `gloss` can contain two (or more) parts conidexed with the individual m-lemmas instead of a single value:

```
<!ELEMENT gloss (#PCDATA|coindexed)>
<!ELEMENT coindexed (#PCDATA)>
<!ATTLIST coindexed
     coindex CDATA #REQUIRED
>
```

The element `example` contains an example usage of the given LU (one or more sentences or sentence fragments):

```
<!ELEMENT example (#PCDATA|coindexed)>
```

The element `control` is present only in LUs that contain infinitive in one of their slots. The content of this element specifies the type of control (see Section 5.9):

```
<!ELEMENT control (#PCDATA)>
```

The element `class` specifies which semantic class the given LU belongs to.[16]

```
<!ELEMENT class (#PCDATA)>
```

Alternations applicable on the given LU are listed in the element named `alternations` (of course, since the alternations in Czech are not sufficiently studied yet, it is not possible to properly specify a list of possible values of the attribute `alt_name` in the presented DTD):

```
<!ELEMENT alternations (alternation+)>
<!ELEMENT alternation (EMPTY)>
<!ATTLIST alternation
     alt_name CDATA #REQUIRED
>
```

Let us focus on the derived LUs now. The element `dlu` is structured in the same way as `blu`, just that the components `gloss` and `example` are not obligatory (derived LUs are supposed to be generated mostly automatically, and it would be extremely difficult to generate also examples of usage and glosses). The element `frame` embedded in `dlu` contains a valency frame resulting from the application of the specified alternation on the frame of the specified input LU:

```
<!ELEMENT dlu (lexical_forms?, frame, gloss?, example?,
               control?, class?, alternations?)>
<!ATTLIST coindexed
     alt_name CDATA #REQUIRED
     input_lu IDREFS #REQUIRED
>
```

Finally, it should be recalled that no alternations have been physically annotated in VALLEX yet and therefore no derived LUs can be generated now. However, the B-version data format itself is prepared for the alternation-based model.

[16]The inventory of verb classes used in VALLEX is still very preliminary and is not mentioned in this thesis.

6.7 Remark on standardization

In the previous sections, several data formats used within the VALLEX project were mentioned. The reader may wonder whether there is any standard for lexicographic data, and if so, then why we did not use it. This is to be answered in this section.

First, there are too many "standards" for lexicographic data, without any of them being dominant, which effectively means that there is no *de facto* standard at all. The following list of standardization initiatives is by far not exhaustive:[17]

- EAGLES (Expert Advisory Group on Language Engineering Standards)[18] and its successor ISLE (International Standards for Language Engineering)[19]
- PAROLE[20] and its continuation SIMPLE[21]
- OLIF (Open Lexicon Interchange Format)[22]
- SALR (Standards-based Access service to multilingual Lexicons and Terminologies)[23]
- CONCEDE (CONsortium for Central European Dictionary Encoding)[24]

Second and more importantly, none of the standards we have studied so far matches our needs: either the special VALLEX structure violates the expectations of the creators of the given standard (i.e., the standard cannot be monotonically arranged to match the VALLEX properties), or it is potentially convertible into the standard, but just because of the fact that the standard requirements are very loose. For instance, VALLEX could be easily transformed into a feature-structure format, preliminarily suggested by a joint initiative of the TEI and ISO Committee TC37SC 4 ([Lee et al., 2004]), but – in our opinion – saying that a discrete linguistic data structure can be represented as a general feature structure is not much more informative than saying that it can be represented as a sequence of ones and zeros. The (still doubtable) advantage gained by twisting the data into one of these standard formats would not compensate the burden of technical problems (e.g.

[17] When exploring the existing lexicographic formats, we used an unpublished survey written up by Eduard Bejček.

[18] http://www.ilc.cnr.it/EAGLES96/home.html

[19] http://www.ilc.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm

[20] http://www.ub.es/gilcub/SIMPLE/simple.html

[21] http://www.ub.es/gilcub/SIMPLE/simple.html

[22] http://www.olif.net/

[23] http://www.ttt.org/salt/

[24] http://www.itri.brighton.ac.uk/projects/concede/

cumbersome querying) related to the fact that the physical data structure would not naturally mirror the logical one.

The conclusion is that we are not aware of any available "standard" data format which would suit our lexicon. In our opinion, this is not such a bad news as it seems: thanks to the contemporary general turn to XML, conversion of one (XML-based) lexicon format into another (XML-based) lexicon format is from the technological point of view a much simpler task when compared e.g. to conversion of two different typesetters' tape formats a couple of decades ago.[25] But this is of course not to say that the attempts at further standardization are not creditable.

6.8 Querying VALLEX in XSH

So far, we have mentioned three ways how VALLEX can be searched: (i) one can search the text files in the primary format directly in the text editor (Section 6.1), (ii) one can specify queries based on regular expressions in the WWW search interface (Section 6.3), or (iii) one can find (statically pre-generated) answers to some basic questions in the HTML version of VALLEX (Section 6.5). However, it is clear that none of these environments is suitable for more complex queries, and none of them can be used when the lexicon is to be accessed automatically. In such situations, the XML version of VALLEX has to be used.

Nowadays there is a bunch of software technologies developed for XML, and basically all modern programming languages are equipped with libraries for processing XML data; the choice is only up to the VALLEX user. In this section we briefly demonstrate one of the possible approaches, namely accessing VALLEX via XSH.

XSH[26] is a powerful command-line tool for querying, processing and editing XML documents developed by Petr Pajas, allowing both interactive work and batch processing. It supports working with more than one Document Object Model at once, filesystem-like navigation within the DOM tree using XPath, combining XPath with Perl and shell commands, and many other features. The XSH session with VALLEX can be started as follows:

```
osiris:...txt2newxml/ xsh
------------------------------------------------------------
xsh - XML Editing Shell version 1.8.2/0.12 (Revision:  1.33)
------------------------------------------------------------
Copyright (c) 2002 Petr Pajas.
This is free software, you may use it and distribute it under
```

[25] Supposing that the two XML structures are well designed and reasonably mirror the logical structure of the data.

[26] urlhttp://xsh.sourceforge.net/

```
either the GNU GPL Version 2, or under the Perl Artistic License.
Using terminal type:  Term::ReadLine::Perl
Hint:  Type 'help' or 'help | less' to get more help.
xsh scratch:/> open v=vallex_b_1.5.xml
parsing vallex_b_1.5.xml
done.
xsh v:/>
```

One can enter the root element (as if it was a directory in a filesystem):

```
xsh v:/> cd vallex_b
```

and view its content:

```
xsh v:/vallex_b> ls
<vallex_b>
<head>...</head>
<body>...</body>
</vallex_b>
Found 1 node(s).
```

or enter a more specific element using its XPath position (and print its surrounding):

```
xsh v:/> cd //mlemma[text()='zvolat']
xsh v:/vallex_b/body/entries/lexeme_cluster[1236]/
lexeme/lexical_forms/mlemma> ls ../..
<lexeme pos="v" id="lxm-v-zvolat">
<lexical_forms>
<mlemma aspect="pf" coindex="pf">zvolat</mlemma>
</lexical_forms>
<lexical_units>
<lu_cluster id="luc-v-zvolat-1">
<blu id="blu-v-zvolat-1">
<frame>
<slot functor="ACT" type="obl">
<form type="direct_case" case="1"/>
</slot>
<slot functor="ADDR" type="opt">
<form type="prepos_case" prepos_lemma="na" case="4"/>
</slot>
<slot functor="PAT" type="obl">
<form type="direct_case" case="4" to_be="0"/>
<form type="subord_conj" subord_conj_lemma="jak"/>
<form type="subord_conj" subord_conj_lemma="že"/>
<form type="rel"/>
</slot>
```

```
</frame>
<gloss>výrazně a hlasitě pronést</gloss>
<example>zvolal, že to není možné</example>
<class>communication</class>
</blu>
</lu_cluster>
</lexical_units>
</lexeme>
Found 1 node(s).
```

M-lemmas of irreflexive lexemes containing more than 20 LUs can be printed as follows:

```
xsh v:/> ls //lexeme[count(.//blu)>20 and
count(.//reflex)=0]/lexical_forms//mlemma/text()
brát
brávat
vzít
dávat
dát
padat
padnout
Found 7 node(s).
```

The number of occurrences of various types of lexicon items can be counted easily. The number of frame slots potentially expressed by infinitive is obtained by the following command:

```
xsh v:/> count //slot/form[@type="infinitive"]
303
```

Shell commands can be combined with XPAth, such as in the following example, which prints ten most frequent functors:

```
xsh v:/> ls //@functor | sort | uniq -c | sort -nr | head -n10
Found 12313 node(s).
4403 functor='ACT'
3440 functor='PAT'
520 functor='ADDR'
434 functor='MEANS'
429 functor='DIR3'
423 functor='LOC'
414 functor='EFF'
397 functor='BEN'
296 functor='MANN'
270 functor='DPHR'
```

6.9 Quantitative properties of VALLEX 1.5

In this section we present VALLEX from the quantitative point of view. For this purpose, we use an unpublished version of the lexicon, internally denoted as VALLEX 1.5, converted into the B-version XML. When determining quantitative properties of the lexicon, we continue in illustrating how VALLEX can be processed in XSH.

Numbers of lexeme clusters and lexemes can be computed as follows:

```
xsh v:/>
xsh v:/> count //lexeme_cluster
1239
xsh v:/> count //lexeme
1624
```

However, if one is interested in the corpus coverage of the lexicon, then the number of different m-lemmas is more important (homographs are counted only once in the following query):[27]

```
xsh v:/> ls //mlemma/text() | sort | uniq | wc -l
Found 3440 node(s).
1841
```

As expected, lexemes containing homographs or lemma variants are not very frequent (however, still too frequent to be completely disregarded in the lexicon design):

```
xsh v:/>
count //lexeme[count (./lexical_forms//mlemma[./@homograph])>0]
154
count //lexeme[count (./lexical_forms/mlemma_variants)>0]
75
```

Reflexive and irreflexive lexemes are combined in the same lexeme clusters with the following frequencies:

- irreflexive + reflexive `se` + reflexive `si` in the same lexeme:
  `xsh v:/> count //lexeme_cluster[count(./lexeme)=3]`
  23
- irreflexive + reflexive `se`:
  `xsh v:/> xsh v:/> count //lexeme_cluster[count(./lexeme)=2`

[27] The corpus coverage of the lexicon seems easy to estimate using e.g. the coverage curve on page 85. However, it is necessary to point out that all such estimations are seriously skewed by the most frequent verb `být` (to be). First, it is not trivial to distinguish autosemantic and auxiliary usages of this verb (in the latter case, e.g. in complex future tense, *být* is not supposed to have its own valency, and thus such occurrences have not to do with lexicon entries). Second, the valency of the meaningful *být* is extremely intricate, and its description in VALLEX is still far from stable.

```
and count(./lexeme/lexical_forms/reflex[text()="si"])=0]
298
```

- irreflexive + reflexive `si`:
```
xsh v:/> count //lexeme_cluster[count(./lexeme)=2
and count(./lexeme/lexical_forms/reflex[text()="se"])=0]
36
```
- reflexive `se` + reflexive `si`:
```
xsh v:/> count //lexeme_cluster[count(./lexeme)=2
and count(./lexeme/lexical_forms/reflex[text()="se"])=1
and count(./lexeme/lexical_forms/reflex[text()="si"])=1]
5
```
- irreflexive:
```
xsh v:/> count //lexeme_cluster[count(./lexeme)=1
and count(./lexeme/lexical_forms/reflex)=0]
734
```
- reflexive `se`:
```
xsh v:/> count //lexeme_cluster[count(./lexeme)=1
and count(./lexeme/lexical_forms/reflex[text()="se"])=1]
123
```
- reflexive `si`:
```
xsh v:/> count //lexeme_cluster[count(./lexeme)=1
and count(./lexeme/lexical_forms/reflex[text()="si"])=1]
20
```

Now let us look at perfectives and imperfectives:

- both perfective and imperfective in the same lexeme:[28]
```
xsh v:/> count //lexeme[count(./lexical_forms//mlemma
[@aspect="impf"])>0 and count(./lexical_forms//
mlemma[@aspect="pf"])>0]
905
```
- imperfective only:
```
xsh v:/> count //lexeme[count(./lexical_forms//mlemma
[@aspect="impf"])>0  and count(./lexical_forms//
mlemma[@aspect="pf"])=0]
469
```
- perfective only:
```
xsh v:/> count //lexeme[count(./lexical_forms//mlemma
[@aspect="impf"])=0 and count(./lexical_forms//
mlemma[@aspect="pf"])>0]
212
```

Now we present an empirical observation which supports our decision to

[28] If the size of VALLEX 1.5 is to be compared to 1400 'verbs' contained in VALLEX 1.0, one should count the lexemes containing aspectual pairs twice ($1624 + 905 = 2529$).

merge aspectual counterparts into single lexemes in VALLEX. First we count the total number of LU clusters in lexemes having both perfective and imperfective form:

```
xsh v:/> count //lexeme[count(./lexical_forms//
mlemma[@aspect="impf"])>0 and count(./lexical_forms//
mlemma[@aspect="pf"])>0]/lexical_units/*
2728
```

Second, we count how many times it was necessary to specify lexical forms directly for a lexical unit (which is in VALLEX 1.5 caused almost exclusively by inapplicability of one of the aspectual counterparts):

```
xsh v:/> count //blu/lexical_forms
564
```

As we can see, around 80 % of LUs in lexemes having perfective and imperfective forms are associated with both forms; thus it is more economic to treat the rest as exceptions than to always treat the aspectual counterparts as completely separate lexemes.

The total number of LUs in the lexicon

```
xsh v:/> count //blu
4414
```

gives the average of 2.7 LUs per lexeme. The total number of frame slots

```
xsh v:/> count //slot
12313
```

gives the average of 2.8 slots per frame.

Chapter 7

Final Remarks

> The wrong thing about the dictionaries
> is that people believe they are correct.
> **Werner Lansburgh's friend Johny**

We believe we have achieved three important goals when working on this thesis:

- The primordial aim of the presented work was to create a publicly available high-quality NLP-oriented lexical resource containing valency frames of (a subset of) Czech verbs. This aim was first achieved in the autumn of 2003 when we released the first version of VALLEX (the further development of which still continues). VALLEX 1.0 is freely available for non-commercial purposes; there are more than 100 registered users at this time point. VALLEX as such is a collective work, the present author contributed to its creation especially by implementing the components of the dictionary production system.
- Second, we have collected dispersed linguistic knowledge necessary for building the valency lexicon (paying special attention to the choice of terminology), and have tried to further elaborate the background valency theory adopted from Functional Generative Description. We have explicitly split the description of valency into two levels (surface and deep) and have demonstrated that the relation between the two levels is not trivial. We also have introduced new terminology for describing instances of valency frames in language use. We have sketched an alternation-based lexicon model, which in our opinion provides us with an efficient mechanism for capturing different types of regularities in the lexicon and thus will help to reduce its redundancy (however, this will require basic linguistic research first).
- Third, we hope that the presented survey of existing language resources related to valency can be interesting also for other researchers in the field. We are not aware of any other comparatively wide review on this topic.

The existence of VALLEX also has influenced some other projects:

- Experience from the development of VALLEX (as well as some data containing entries for several hundred verbs) was used in the initial stage of PDT-VALLEX ([Hajič et al., 2003]).
- More recently, some components of the presented dictionary production system (including the annotation format and the idea of using a syntax-highlighting text editor for creating the entries; including the tools for converting the data into XML, HTML, and PDF, as well as the graphical layouts) were adapted within the VerbaLex project presented in [Hlaváčková and Horák, 2005].
- Some principles of VALLEX were also used in the Swedish-Czech lexicon of verbonominal constructions ([Cinková and Žabokrtský, 2005]).
- The existence of VALLEX made it possible to annotate a sample of the Czech National Corpus with valency frames. The resulting resource is called VALEVAL and contains 10,000 corpus instances (100 verbs, 100 sentences per verb), each of them manually assigned with a valency frame from VALLEX and automatically tagged and parsed. VALEVAL was already used for an experiment with an automatic frame assignment ([Bojar et al., 2005]), which is to our knowledge the first experiment on Word Sense Disambiguation based on large data in Czech, and which becomes the first tangible NLP application of VALLEX.

Bibliography

[Allen, 1995] Allen, J. (1995). *Natural Language Understanding*. Benjamin/Cummings Publishing Company. 2nd edition.

[Babby, 1998] Babby, L. H. (1998). Voice and Diathesis in Slavic. Workshop on Slavic Morphosyntax: State of the art.

[Babko-Malaya et al., 2004] Babko-Malaya, O., Palmer, M., Xue, N., Joshi, A., and Kulick, S. (2004). Proposition Bank II: Delving Deeper. In Meyers, A., editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 17–23, Boston, Massachusetts, USA. Association for Computational Linguistics.

[Balabanova and Ivanova, 2002] Balabanova, E. and Ivanova, K. (2002). Creating a machine-readable version of bulgarian valence dictionary (a case study of clark system application). In *Proceedings of The First Workshop on Treebanks and Linguistic Theories*, Sozopol, Bulgaria.

[Bémová, 1979] Bémová, A. (1979). Synktaktické vlastnosti prefigovaných sloves. *Explicite Beschreibung der Sprache und automatische Textbearbeitung*, (V. Transducing components of functional generative description 2).

[Blatná et al., 2004] Blatná, R., Čermák, F., Hlaváčová, J., Hnátková, M., Kocek, J., Kopřivová, M., Křen, M., Petkevič, V., Schmiedtová, V., Stluka, M., and Šulc, M. (2004). *Frekvenční slovník češtiny*.

[Boguslavsky et al., 2004] Boguslavsky, I., Iomdin, L., and Sizov, V. (2004). Multilinguality in ETAP-3: Reuse of Lexical Resources. In *Proceedings of PostCOLING Workshop on Multilingual Linguistic Resources*.

[Bojar et al., 2005] Bojar, O., Semecký, J., and Benešová, V. (2005). VALEVAL: Testing VALLEX Consistency and Experimenting with Word-Frame Disambiguation. *Prague Bulletin of Mathematical Linguistics*, 83.

[Bolshakov and Gelbukh, 2000] Bolshakov, I. A. and Gelbukh, A. F. (2000). The Meaning-Text Model: Thirty Years After. International Forum on Information and Documentation.

[Bond and Shirai, 1997] Bond, F. and Shirai, S. (1997). Practical and Efficient Organization of a Large Valency Dictionary. In *Proceedings of the 4th Natural Language Processing Pacific*, Phuket, Thailand.

[Brants et al., 2002] Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The tiger treebank. In *In Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol, Bulgaria.

[Briscoe, 2001] Briscoe, T. (2001). From dictionary to corpus to self-organizing dictionary: learning valency associations in the face of variation and change. In Rayson, P., Wilson, A., McEnery, T., Hardie, A., and Khoja, S., editors, *Proceedings of Corpus Linguistics 2001*, pages 79–89.

[Burchardt et al., 2005] Burchardt, A., Frank, A., and Pinkal, M. (2005). Building Text Meaning Representations from Contextually Related Frames – A Case Study. In *Proceedings of 6th International Workshop on Computational Semantics*, Tilburg, The Netherlands.

[Bybee, 1985] Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form* . Benjamins, Philadelphia.

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

[Cinková and Žabokrtský, 2005] Cinková, S. and Žabokrtský, Z. (2005). Treating support verb constructions in a lexicon: Swedish-Czech combinatorial valency lexicon of predicate nouns. In *Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, pages 28–31. Saarland University.

[Civit et al., 2005] Civit, M., Marti, M. A., Morante, R., Navarro, B., Taule, M., and Aldezabal, I. (2005). Defining a framework for the analysis of predicates. In *Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, pages 28–31. Saarland University.

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

[Daneš, 1994] Daneš, F. (1994). The Sentence-Pattern Model of Syntax. In Luelsdorff, P. A., editor, *The Prague School of Structural and Functional Linguistics*, pages 197–221. John Benjamins Publishing Company.

[Daneš and Hlavsa, 1987] Daneš, F. and Hlavsa, Z. (1987). *Větné vzorce v češtině*. Academia, Praha.

[Davis, 2001] Davis, A. R. (2001). *Linking by Types in the Hierarchical Lexicon*. CSLI Publications.

[den Eynde and Blanche-Benveniste, 1978] den Eynde, K. V. and Blanche-Benveniste, C. (1978). Syntaxe et mécanismes descriptifs: présentation de l'approche pronominale cahiers de lexicologie 32, 3-27.

[Dowty, 1986] Dowty, D. (1986). On the Semantic Content of the Notion Thematic Role. In *Property Theory, Type Theory and Natural Language Semantics*. Dordrecht: Reidel.

[Eisner, 2002] Eisner, P. (2002). *Rady Čechům, jak se hravě přiučiti češtině*. Academia.

[Ellsworth et al., 2004] Ellsworth, M., Erk, K., Kingsbury, P., and Pado, S. (2004). PropBank, SALSA, and FrameNet: How design determines product. In *Proceedings of the LREC 2004 Workshop on Building Lexical Resources from Semantically Annotated Corpora*, Lisbon.

[Erk et al., 2003] Erk, K., Kowalski, A., Pado, S., and Pinkal, M. (2003). Towards a Resource for Lexical Semantics: A Large German Corpus with Extensive Semantic Annotation. In *Proceedings of ACL-03*, Sapporo, Japan.

[Fellbaum and Miller, 2003] Fellbaum, C. and Miller, G. A. (2003). Morphosemantic Links in WordNet. *Traitement automatique des languages*, 44(2):69–80.

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

[Fillmore, 1968] Fillmore, C. J. (1968). The case for case. In Bach, E. and Harms, R., editors, *Universals in Linguistic Theory*, pages 1–90. New York.

[Fillmore, 2002] Fillmore, C. J. (2002). FrameNet and the Linking between Semantic and Syntactic Relations. In Tseng, S.-C., editor, *Proceedings of COLING 2002*, pages xxviii–xxxvi. Howard International House.

[Fillmore et al., 2002] Fillmore, C. J., Baker, C., and Sato, H. (2002). Seeing Argumens through Transparent Structures. In Rodríguez, M. G. and Araujo, C. P. S., editors, *Proceedings of LREC 2002*, pages 787–791. ELRA.

[Fujita and Bond, 2002] Fujita, S. and Bond, F. (2002). Extending the Coverage of a Valency Dictionary. In *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, Taiwan.

[Grepl and Karlík, 1986] Grepl, M. and Karlík, P. (1986). *Skladba spisovné češtiny*. Státní pedagogické nakladatelství Praha.

[Hajič et al., 2003] Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V., and Pajas, P. (2003). PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. volume 9 of *Mathematical Modeling in Physics, Engineering and Cognitive Sciences*, pages 57–68.

[Hajičová and Kučerová, 2002] Hajičová, E. and Kučerová, I. (2002). Argument/Valency Structure in PropBank, LCS Database and Prague Dependency Treebank: A Comparative Pilot Study. pages 846–851.

[Hajičová and Sgall, 2003] Hajičová, E. and Sgall, P. (2003). Dependency syntax in Functional Generative Description. In *Dependenz und Valenz – Dependency and Valency*, volume I, pages 570–592. Walter de Gruyter. MSM113200006, LN00A063.

[Helbig and Schenkel, 1969] Helbig, G. and Schenkel, W. (1969). *Wörterbuch zur Valenz und Distribution deutscher Verben*. VEB BIBLIOGRAPHISCHES INSTITUT, Leipzig, Germany.

[Hlaváčková and Horák, 2005] Hlaváčková, D. and Horák, A. (2005). Transformation of WordNet Czech Valency Frames into Augmentech VALLEX-1.0 Format. In *Proceedings of the 2nd Language & Technology Conference*, Poznaň.

[Hnátková, 2002] Hnátková, M. (2002). Značkování frazémů a idiomů v Českém národním korpusu s pomocí slovníku české frazeologie a idiomatiky. *Slovo a slovesnost*, (63):117–126.

[Horák, 1998] Horák, A. (1998). Verb valency and semantic classification of verbs. In Sojka, P., Matoušek, V., Pala, K., and Kopeček, I., editors, *Text, Speech and Dialogue - TSD 98*, Brno.

[Kahane, 2003] Kahane, S. (2003). The meaning-text theory. *Dependency and Valency. An International Handbook of Contemporary Research*.

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

[Katz and Fodor, 1963] Katz, S. M. and Fodor, J. A. (1963). The structure of a semantic theory. *Language*, 39:170–200.

[Kilgarriff, 1992] Kilgarriff, A. (1992). *Polysemy*. PhD thesis, University of Sussex.

[Kingsbury and Palmer, 2002] Kingsbury, P. and Palmer, M. (2002). From Treebank to PropBank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Las Palmas, Spain.

[Kingsbury and Palmer, 2003] Kingsbury, P. and Palmer, M. (2003). PropBank–the Next Level of TreeBank. In Nivre, J. and Hinrichs, E., editors, *In proceedings of the Second Workshop on Treebanks and Linguistic Theories*, pages 105–116, Växjö, Sweden. Växjö University Press.

[Kingsbury et al., 2002] Kingsbury, P., Palmer, M., and Marcus, M. (2002). Adding Semantic Annotation to the Penn TreeBank. In *Proceedings of the Human Language Technology Conference*, San Diego, California.

[Kopečný, 1962] Kopečný, F. (1962). *Slovesný vid v češtině*. Nakladatelství Československé akademie věd.

[Koster and Gradmann, 2004] Koster, C. H. and Gradmann, S. (2004). The language belongs to the people! In *Proceedings of Fourth International Conference on Language Resources and Evaluation, LREC 2004*, pages 353–356.

[Kunze and Rösner, 2004] Kunze, M. and Rösner, D. (2004). Corpus based Enrichment of GermaNet Verb Frames. In *Proceedings of Fourth International Conference on Language Resources and Evaluation, LREC 2004*, pages 963–966.

[Kučová et al., 2003] Kučová, L., Kolářová, V., Žabokrtský, Z., Pajas, P., and Čulo, O. (2003). Anotování koreference v Pražském závislostním korpusu. Technical report.

[Kučová and Žabokrtský, 2005] Kučová, L. and Žabokrtský, Z. (2005). Anaphora in Czech: Large Data and Experiments with Automatic Anaphora Resolution . In *Proceedings of 8th International Conference on Text, Speech and Dialogue*.

[Lee et al., 2004] Lee, K., Burnard, L., Romary, L., de la Clergerie, E., Declerck, T., Bauman, S., Bunt, H., Clément, L., Erjavec, T., Roussanaly, A., and Roux, C. (2004). Towards and international standard on feature structure representation. In *Proceedings of Fourth International Conference on Language Resources and Evaluation, LREC 2004*, pages 373–376.

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

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

[Lopatková et al., 2002] Lopatková, M., Řezníčková, V., and Žabokrtský, Z. (2002). Valency Lexicon for Czech: from Verbs to Nouns. In *Proceedings of 5th International Conference on Text, Speech and Dialogue*, volume 2448 of *Lecture Notes in Artificial Intelligence*, pages 147–150. LN00A063.

[Lopatková et al., 2003] Lopatková, M., Žabokrtský, Z., Skwarska, K., and Benešová, V. (2003). VALLEX 1.0 Valency Lexicon of Czech Verbs. Technical Report TR-2003-18.

[Manning and Sag, 1998] Manning, C. D. and Sag, I. A. (1998). Argument structure, valence, and binding. *Nordic Journal of Linguistics*, 21.

[Mel'čuk, 2004] Mel'čuk, I. (2004). Actants in semantics and syntax I: actants in semantics. *Linguistics*, 42 (1):1–66.

[Mel'čuk, 1988] Mel'čuk, I. A. (1988). *Dependency Syntax: Theory and Practice*. State University of New York Press.

[Mel'čuk and Zholkovsky, 1984] Mel'čuk, I. A. and Zholkovsky, A. K. (1984). *Explanatory Combinatorial Dictionary of Modern Russian*. Wiener Slawistischer Almanach, Vienna.

[Meyers et al., 2004] Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., and Grishman, R. (2004). The NomBank Project: An Interim Report. In Meyers, A., editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31, Boston, Massachusetts, USA. Association for Computational Linguistics.

[Miltsakaki et al., 2004] Miltsakaki, E., Joshi, A., Prasad, R., and Webber, B. (2004). Annotating Discourse Connectives and Their Arguments. In Meyers, A., editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 9–16, Boston, Massachusetts, USA. Association for Computational Linguistics.

[Nasr and Rambow, 2004] Nasr, A. and Rambow, O. (2004). SuperTagging and Full Parsing. In *Proceedings of Seventh International Workshop on Tree Adjoining Grammar and Related Formalisms*.

[Nižníková and Sokolová, 1998] Nižníková, J. and Sokolová, M. (1998). *Valenčný slovník slovenských slovies*. Filozofická fakulta Prešovskej univerzity.

[Pala and Smrž, 2004] Pala, K. and Smrž, P. (2004). Building czech wordnet. *ROMANIAN JOURNAL OF INFORMATION, SCIENCE AND TECHNOLOGY*, 7(1-2):79–88.

[Pala and Ševeček, 1997] Pala, K. and Ševeček, P. (1997). Valence českých sloves. In *Sborník prací FFBU*, pages 41–54, Brno.

[Panevová, 1974] Panevová, J. (1974). On Verbal Frames in Functional Generative Description. In *The Prague Bulletin of Mathematical Linguistics 22*, pages 3–40.

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

[Panevová, 1996] Panevová, J. (1996). More Remarks on Control. *Prague Linguistic Circle Papers, John Benjamins*, 2:101–120.

[Panevová et al., 1971] Panevová, J., Benešová, E., and Sgall, P. (1971). *Čas a modalita v češtině*. Universita Karlova, Praha.

[Pauliny, 1943] Pauliny, E. (1943). *Štruktúra slovenského slovesa*. Bratislava.

[Polański, 1992] Polański, K., editor (1980-1992). *Słownik syntaktyczno-generatywny czasowników polskich*. Wydawnictwo Polskiej Akademii Nauk, Wrocław.

[Popova, 1987] Popova, M. (1987). *Kratak valenten rechnik na glagolite v savremennia bulgarski knizoven ezik*. Bulgarian Academy of Sciences Publishing House.

[Pustejovsky, 1995] Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press, Cambridge.

[Pustejovsky et al., 2004] Pustejovsky, J., Hanks, P., and Rumshisky, A. (2004). Automated Induction of Sense in Context. In Hansen-Schirra, S., Oepen, S., and Uszkoreit, H., editors, *COLING 2004 5th International Workshop on Linguistically Interpreted Corpora*, pages 55–58, Geneva, Switzerland. COLING.

[Rambow et al., 2003] Rambow, O., Dorr, B., Kipper, K., Kučerová, I., and Palmer, M. (2003). Automatically Deriving Tectogrammatical Labels from Other Resources: A Comparison of Semantic Labels Across Frameworks. *Prague Bulletin of Mathematical Linguistics*, (79–80):23–35. ME642, LN00A063.

[Razímová and Žabokrtský, 2005] Razímová, M. and Žabokrtský, Z. (2005). Morphological Meanings in the Prague Dependency Treebank 2.0. In *to appear in the proceedings of Text, Speech and Dialogue 2005*.

[Resnik, 1993] Resnik, P. (1993). *Selection and Information: A Class-Based Approach to Lexical Relationships*. PhD thesis, University of Pennsylvania.

[Sgall, 1967] Sgall, P. (1967). *Generativní popis jazyka a česká deklinace*. Academia.

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

[Sgall et al., 1986] Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company, Dordrecht.

[Silnickij, 1999] Silnickij, G. (1999). *Korreljacionnaja tipologija glagolnych sistem indoevropejskich i inostrukturnych jazykov*. Smolensk, Russia.

[Skoumalová, 2001] Skoumalová, H. (2001). *Czech syntactic lexicon*. PhD thesis, Univerzita Karlova, Filozofická fakulta.

[Skoumalová, 2002] Skoumalová, H. (2002). Verb frames extracted from dictionaries. *The Prague Bulletin of Mathematical Linguistics 77*.

[Somers, 1986] Somers, H. L. (1986). The need for mt-oriented versions of case and valency in mt. In *COLING*, pages 118–123.

[Součková, 2005] Součková, K. (2005). Valence sloves mluvení. Master's thesis, Faculty of Philosophy and Art, Charles University in Prague.

[SSJČ, 1978] SSJČ (1978). *Slovník spisovné češtiny pro školu a veřejnost*. Academia, Praha.

[Stevenson, 2003] Stevenson, M. (2003). *Word Sense Disambiguation: The Case for Combinations of Knowledge Sources*. CSLI Publications.

[SČFI, 1983] SČFI (1983). *Slovník české frazeologie a idiomatiky*. Academia, Praha.

[Svozilová et al., 1997] Svozilová, N., Prouzová, H., and Jirsová, A. (1997). *Slovesa pro praxi*. Academia, Praha.

[Tabakowska, 2003] Tabakowska, E. (2003). Those notorious polish reflexive pronouns: a plea for middle voice. http://www.seelrc.org/glossos/issues/4/tabakowska.pdf.

[Tesnière, 1959] Tesnière, L. (1959). *Eléments de syntaxe structurale*. Paris.

[Uher, 1987] Uher, F. (1987). *Slovesné předpony*. Univerzita J. E. Purkyně, Brno.

[Urešová, 2004] Urešová, Z. (2004). The verbal valency in the Prague Dependency Treebank from the annotator's point of view.

[van den Eynde and Mertens, 2003] van den Eynde, K. and Mertens, P. (2003). La valence: l'approche pronominale et son application au lexique verbal. *French Language Studies*, pages 63–104.

[Verspoor, 1997] Verspoor, C. M. (1997). *Contextually-Dependent Lexical Semantics*. PhD thesis, The University of Edinburgh.

[Vossen, 1998] Vossen, P., editor (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht.

[Žabokrtský, 2005] Žabokrtský, Z. (2005). Resemblances between Meaning-Text Theory and Functional Generative Description. In *Proceedings of Second International Conference on Meaning-Text Theory*, Moscow.

[Wanner, 1996] Wanner, L., editor (1996). *Lexical Functions in Lexicography and Natural Language Processing*. John Benjamins Publishing Company.

Appendix A

Functors used in VALLEX

- ACMP (accompaniement): *Mother came with her children.*
- ACT (actor): *Peter read a letter.*
- ADDR (addressee): *Peter gave Mary a book.*
- AIM (aim): *John came to a bakery for a piece of bread.*
- BEN (benefactive): *She made this for her children.*
- CAUS (cause): *She did so since they wanted it.*
- COMPL (complement): *They painted the wall blue.*
- DIFF (difference): *The number has swollen by 200.*
- DIR1 (direction-from): *He went from the forest to the village.*
- DIR2 (direction-through): *He went through the forest to the village.*
- DIR3 (direction-to): *He went from the forest to the village.*
- DPHR (dependent part of a phraseme): *Peter talked horse again.*
- EFF (effect): *We made her the secretary.*
- EXT (extent): *The temperatures reached an all time high.*
- HER (heritage): *He named the new villa after his wife.*
- INTT (intent): *He came there to look for Jane.*
- LOC (locative): *He was born in Italy.*
- MANN (manner): *They did it quickly.*
- MEANS (means): *He wrote it by hand.*
- NORM (norm): *Peter has to do it exactly according to directions.*
- OBST(obstacle): *The boy stumbled over a stumb.*
- ORIG (origin): *She made a cake from apples.*
- PAT (patient): *I saw him.*
- RCMP (recompense): *She bought a new shirt for 25 $.*
- REG (regard): *With regard to George she asked his teacher for advice.*
- RESL (result): *Mother protects her children from any danger.*
- SUBS (substitution): *He went to the theatre instead of his ill sister.*
- TFHL (temporal-for-how-long): *They interrupted their studies for a year.*
- TFRWH (temporal-from-when): *His bad reminiscences came from this period.*

- THL (temporal-how-long): *We were there for three weeks.*
- TOWH (temporal-to when): *He put it over to next Tuesday.*
- TSIN (temporal-since-when): *I have not heard about him since that time.*
- TWHEN (temporal-when): *His son was born last year.*

Appendix B

VALLEX 1.0 Document Type Definition

```
<!ELEMENT vallex (word_entry+)>

<!ELEMENT word_entry ((headword_lemma|headword_variants),frame_entry+)>
<!ATTLIST word_entry
        aspect (pf|impf|biasp) #REQUIRED
>

<!ELEMENT headword_lemma (#PCDATA)>
<!ATTLIST headword_lemma
        homonym_index CDATA #IMPLIED
>

<!ELEMENT headword_variants (headword_lemma, headword_lemma+)>

<!ELEMENT frame_entry (aspectual_counterparts?, gloss, example,
          control?, class?, frame_slots)>
<!ATTLIST frame_entry
        frame_index CDATA #IMPLIED
        idiom (NO|YES) "NO"
>

<!ELEMENT aspectual_counterparts
        ((counterpart_lemma|counterpart_variants)+)>

<!ELEMENT counterpart_lemma (#PCDATA)>
<!ATTLIST counterpart_lemma
        aspect (pf|impf|biasp|iter) #REQUIRED
        homonym_index CDATA #IMPLIED
        frame_index CDATA #IMPLIED
      >

<!ELEMENT counterpart_variants (counterpart_lemma,counterpart_lemma+)>

<!ELEMENT gloss (#PCDATA)>
<!ELEMENT example (#PCDATA)>
<!ELEMENT control (#PCDATA)>
<!ELEMENT class (#PCDATA)>

<!ELEMENT frame_slots (slot*)>
```

```
<!ELEMENT slot (form*)>
<!ATTLIST slot
   type (obl|opt|typ) #REQUIRED
   functor  (ACT|PAT|ADDR|EFF|ORIG|ACMP|ADVS|AIM|APP|APPS|
      ATT|BEN|CAUS|CNCS|COMPL|COND|CONJ|CONFR|CPR|CRIT|CSQ|
      CTERF|DENOM|DES|DIFF|DIR1|DIR2|DIR3|DISJ|DPHR|ETHD|
      EXT|FPHR|GRAD|HER|ID|INTF|INTT|LOC|MANN|MAT|MEANS|MOD|
      NA|NORM|PAR|PARTL|PN|PREC|PRED|REAS|REG|RESL|RESTR|RHEM|
      RSTR|SUBS|TFHL|TFRWH|THL|THO|TOWH|TPAR|TSIN|TTILL|TWHEN|
      VOC|VOCAT|SENT|DIR|OBST|RCMP)   #REQUIRED
       abbrev (0|1) "0"
     >

<!ELEMENT form EMPTY>
<!ATTLIST form
  type (direct_case|prepos_case|subord_conj|infinitive|
        adjective|phraseme_part) #REQUIRED
  case (1|2|3|4|5|6|7) #IMPLIED
  prepos_lemma (bez|do|jako|k|kolem|kvůli|mezi|místo|na|
       nad|na_úkor|o|od|ohledně|okolo|oproti|po|pod|podle|
       pro|proti|před|přes|při|s|u|v|
       ve_prospěch|vůči|v_zájmu|z|za|než) #IMPLIED
  subord_conj_lemma (že|aby|zda|ať|jak|až) #IMPLIED
  phraseme_part CDATA #IMPLIED
  to_be (0|1) "0"
>
```