Valency Lexicon of Czech Verbs

VALLEX 2.0

Markéta Lopatková, Zdeněk Žabokrtský, Václava Benešová

in cooperation with

Karolína Skwarska, Klára Hrstková, Michaela Nová,
Eduard Bejček, Miroslav Tichý

ÚFAL Technical Report TR-2006-34

ISSN 1214-5521
December 2006

Contents

I VALLEX DESCRIPTION v

1 Introduction vii

2 Logical Structure of the VALLEX Data ix

2.1 Lexemes ix

2.2 Lexical Forms and Lemmas x

2.2.1 Reflexive Lemmas xi

2.2.2 Aspectual Counterparts xi

2.2.3 Lemma Variants xii

2.2.4 Homographs xiii

2.3 Lexical Units xiii

2.4 Valency Frames xv

2.4.1 Functors xvi

2.4.2 Morphemic Forms xvii

2.4.3 Types of Complementations xx

2.4.4 Slot Expansion xx

2.5 Optional LU Attributes xx

2.5.1 Idioms xx

2.5.2 Control xxi

2.5.3 Reflexivity xxi

2.5.4 Reciprocity xxii

2.5.5 Class xxii

II LEXEME ENTRIES xxvii

Part I

VALLEX DESCRIPTION

Chapter 1

Introduction

The Valency Lexicon of Czech Verbs, Version 2.0 (VALLEX 2.0) is a collection of linguistically annotated data and documentation, resulting from an attempt at a formal description of the valency frames of Czech verbs. VALLEX has been developed at the Institute of Formal and Applied Linguistics (ÚFAL) at Faculty of Mathematics and Physics, Charles University in Prague. VALLEX 2.0 is a successor of VALLEX 1.0 ([7]), extended in both theoretical and quantitative aspects.

VALLEX 2.0 provides information on the valency structure of verbs in their particular meanings / senses, possible morphological forms of their complementations and additional syntactic information, accompanied with glosses and examples. All lexeme entries in VALLEX are created manually; manual annotation with accent on consistency is highly time consuming and limit the speed of quantitative growth, but allow for reaching the desired quality.

VALLEX is closely related to the Prague Dependency Treebank (PDT) project. The Functional Generative Description (FGD), being developed by Petr Sgall and his collaborators since the 1960s ([16], [3], [9]), is used as the background theory both in PDT and in VALLEX. In PDT, FGD is being verified by a complex annotation of large amounts of textual data, whereas in VALLEX it is used only for the description of the valency frames of selected verbs.

In VALLEX 2.0, there are roughly 2,730 lexeme entries containing together around 6,460 lexical units (‘senses’). It is important to mention that VALLEX 2.0 – according to FGD and unlike traditional dictionaries – treats a pair of perfective and imperfective aspectual counterparts as a single lexeme. Therefore, if perfective and imperfective verbs are counted separately, the size of VALLEX 2.0 virtually grows to 4,250 entries (still without counting iteratives).

The verbs contained in VALLEX 2.0 were selected as follows: (1) We gradually processed around 2500 most frequent Czech verbs, according to the number of their occurrences in a part of the Czech National Corpus.[1] (2) Simultaneously, we added their perfective or imperfective aspectual counterparts (if they were not already present in the list of the most frequent verbs), and occasionally also iterative counterparts.

[1] http://ucnk.ff.cuni.cz

The preparation of the presented version of VALLEX has taken more than five years. Although it is still the work in progress requiring further linguistic research, we believe that already now it can be useful or at least interesting for other researchers in the field.

Besides this printed version, VALLEX 2.0 is issued also in an electronic form available on the Internet at http://ufal.mff.cuni.cz/vallex/2.0

From the very beginning, VALLEX has been designed with emphasis on both human and machine readability. Therefore, both linguists and developers of applications within the Natural Language Processing domain can use and critically evaluate its content (of course, any feedback from them will be a valuable source of information to us, as well as a great motivation for further work). In order to satisfy different needs of these different potential users, VALLEX 2.0 contains the data in the following three formats:

- Browsable version. The HTML version of the data allows for an easy and fast navigation through the lexicon. Lexemes and lexical units are organized in several ways, following various criteria.
- Printable version. It is identical with Part II of this technical report.
- XML version. Programmers can run sophisticated queries (e.g. based on the XPATH query language) on this machine-tractable data, or use it in their applications.

VALLEX 2.0 has been carried out under the project of the Ministry of Education, Youth and Sports of the Czech Republic No. MSM0021620838 (Objects of Research). It was partially supported also by the projects of the Grant Agency of the Czech Republic No. 405/04/0243, the project of the Ministry of Education, Youth and Sports of the Czech Republic No. LC 536, and the project of the Academy of Science of the Czech Republic, Information Society No, 1ET100300517.

When creating VALLEX 2.0, we have used the following Czech dictionaries (some of them via the dictionary browser DEBDict[2]):

- BRIEF [10],
- Slovník spisovné češtiny [19],
- Slovník spisovného jazyka českého [18],
- Slovesa pro praxi [20],
- Slovník slovesných, substantivních a adjektivních vazeb a spojení [21].

Many thanks for an extensive linguistic and also technical advice go to our colleagues from ÚFAL, especially to professor Jarmila Panevová.

[2] http://nlp.fi.muni.cz/projekty/deb2/debdict/index.html

Chapter 2

Logical Structure of the VALLEX Data

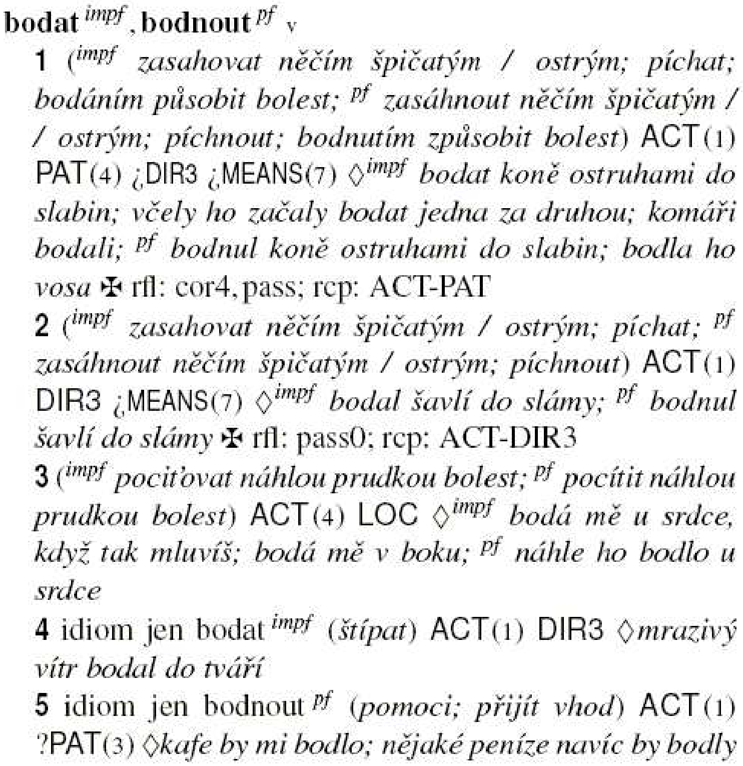
The primary goal of the following text is to briefly describe the content of VALLEX 2.0 data from a structural point of view. Linguistic issues requiring a extensive explanation or discussion are mostly left apart. However, more detailed description (and also additional relevant references) can be found in [22].

The description of the VALLEX 2.0 structure is slightly simplified here, in order to correspond straightforwardly to the visual form of the lexicon and to be sufficient for its full understanding. (The graphical layout of the lexeme entries in the printed version of the lexicon is illustrated in Figure 2.1.) However, it neglects certain features present in the underlying XML version of the lexicon, from which both the printed and html version have been generated. Again, the details about the (slightly richer) XML structure are available in [22].

As for terminology, the terms used here either belong to the broadly accepted linguistic terminology, or come from the Functional Generative Description (which we have used as the background theory), or are defined somewhere else in this text.

2.1 Lexemes

On the highest level, VALLEX 2.0 is composed of **lexemes**. We understand a lexeme as a two-fold abstract entity: it associates a set of possible **lexical forms** (by which the presence of the lexeme is manifested in an utterance) with a set of **lexical units** (complexes of syntactic and semantic features, LUs for short). In simpler words, lexical forms can be viewed as the conjugated forms of a given verbal lexeme, whereas each LU corresponds roughly to the lexeme used in a specific sense and with specific syntactic combinatorial potential. This view is illustrated in Figure 2.2.



bodat^{impf}, **bodnout**^{pf} v

1 (^{impf} *zasahovat něčím špičatým / ostrým; píchat; bodáním působit bolest;* ^{pf} *zasáhnout něčím špičatým / / ostrým; píchnout; bodnutím způsobit bolest*) ACT(1) PAT(4) ¿DIR3 ¿MEANS(7) ◊^{impf} *bodat koně ostruhami do slabin; včely ho začaly bodat jedna za druhou; komáři bodali;* ^{pf} *bodnul koně ostruhami do slabin; bodla ho vosa* ✠ rfl: cor4, pass; rcp: ACT-PAT

2 (^{impf} *zasahovat něčím špičatým / ostrým; píchat;* ^{pf} *zasáhnout něčím špičatým / ostrým; píchnout*) ACT(1) DIR3 ¿MEANS(7) ◊^{impf} *bodal šavlí do slámy;* ^{pf} *bodnul šavlí do slámy* ✠ rfl: pass0; rcp: ACT-DIR3

3 (^{impf} *pociťovat náhlou prudkou bolest;* ^{pf} *pocítit náhlou prudkou bolest*) ACT(4) LOC ◊^{impf} *bodá mě u srdce, když tak mluvíš; bodá mě v boku;* ^{pf} *náhle ho bodlo u srdce*

4 idiom jen bodat^{impf} (*štípat*) ACT(1) DIR3 ◊*mrazivý vítr bodal do tváří*

5 idiom jen bodnout^{pf} (*pomoci; přijít vhod*) ACT(1) ?PAT(3) ◊*kafe by mi bodlo; nějaké peníze navíc by bodly*

Figure 2.1: Sample of a lexeme in a printed form.

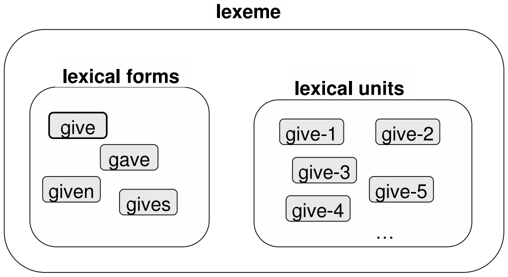


Figure 2.2: Illustration of the notions of *lexeme*, *lexical form*, and *lexical unit*.

2.2 Lexical Forms and Lemmas

It is usual in dictionaries, that the set of all possible lexical forms of a given lexeme is represented only by the infinitive form called lemma.

Lemma in VALLEX 2.0 should be considered as a complex structure:

- it always contains the ‘base’ infinitive form,
- it is always labeled in superscript with its morphological aspect (Section 2.2.2),
- it may contain also reflexive particle (e.g. *bát se*, see Section 2.2.1),
- it may be also labeled with a Roman number in subscript if it is necessary to dis-

tinguish it from its homograph (e.g. *nakupovat*$_I$ - to buy vs. *nakupovat*$_{II}$ - to heap, see Section 2.2.4).

In VALLEX 2.0, there are typically two or more lemmas listed at the beginning of the lexeme entry. It follows FGD principle of treating aspectual counterparts (perfective and imperfective verbs expressing the same lexical meaning, Section 2.2.2) as manifestations of the same lexeme. Another reason for more lemmas being present in the same lexeme might be the existence of orthographic variants (Section 2.2.3).

By default, a LU 'inherits' all lemmas specified for the given lexeme in which it is embedded in VALLEX 2.0. However, it might happen that for a given LU not all the forms specified for the whole lexeme are applicable. In such cases, the list of applicable lemmas is specified for the given LU separately, as in the case of the 4th or 5th LU in Figure 2.1.

2.2.1 Reflexive Lemmas

In VALLEX 2.0, two types of reflexive constructions are distinguished:

- Reflexive lexemes – both true reflexives (e.g. *bát se*, *smát se*) and derived reflexives (e.g. *odpovídat se*, *šířit se*, *vrátit se*) are represented as separate lexemes, and the reflexive particles *se* or *si* are considered as parts of their lemmas.
- Reflexive usage of irreflexive lexemes – if the reflexive particles/pronouns *se* or *si* have specific syntactic function(s), reflexive forms of particular verbs are treated within irreflexive lexemes and their possible functions are specified (see Sections 2.5.3 and 2.5.4) - *se* or *si* can be the part of the reflexive passive form, (e.g., in *pátrá se po zloději*); it can be the complementation coreferential with ACTor (e.g., *mýt se*), or it can mark reciprocity (e.g., *kopat se* in *kopou se vzájemně do nohou*).

2.2.2 Aspectual Counterparts

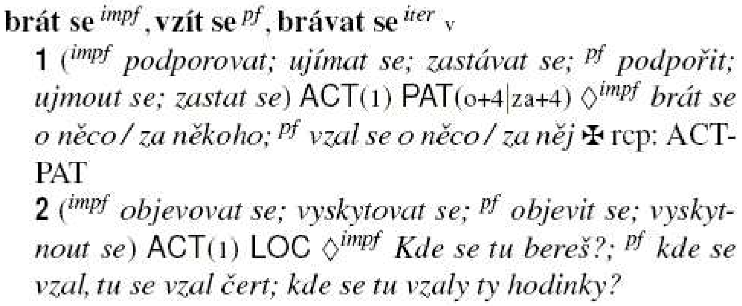
Imperfective and perfective verb forms are distinguished in Czech (as well as a specific subclasses of iterative verbs and of so called biaspectual verbs); this characteristic is called aspect.

In VALLEX 2.0, the value of aspect is attached to each lemma as a superscript label:

- *impf* for imperfective,
- *pf* for perfective,
- *iter* for iterative verbs,
- *biasp* for biaspectual verbs.

There are three ways how aspectual counterparts (verbs with the same or very similar lexical meaning differing in aspect) are formed in Czech (sorted according to productivity):

- *affixation*: imperfective verb is derived from the perfective one, e.g. by infix *-ova-*: *vypsat* / *vypisovat* (to excerpt, to write off);



brát se ^{impf}, **vzít se** ^{pf}, **brávat se** ^{iter} v

1 (^{impf} *podporovat; ujímat se; zastávat se;* ^{pf} *podpořit; ujmout se; zastat se*) ACT(1) PAT(o+4|za+4) ◊^{impf} *brát se o něco / za někoho;* ^{pf} *vzal se o něco / za něj* ✠ rcp: ACT-PAT

2 (^{impf} *objevovat se; vyskytovat se;* ^{pf} *objevit se; vyskytnout se*) ACT(1) LOC ◊^{impf} *Kde se tu bereš?;* ^{pf} *kde se vzal, tu se vzal čert; kde se tu vzaly ty hodinky?*

Figure 2.3: Sample of a lexeme containing both perfective and imperfective verbs.

- *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).

Aspectual counterparts of the first and third type constitute a single lexeme in VALLEX 2.0, as e.g. in the case of *nasedat*^{impf}, *nasednout*^{pf}, *nasedávat*^{iter} (see also Figure 2.3)).

As already mentioned, a LU typically shares all its lemmas with the other LUs in the lexeme in which it is embedded. However, there are exceptions: the aspectual counterpart(s) need not be the same for all LUs of the particular lexeme. For example, *odpovědět*^{pf} is a counterpart of *odpovídat*^{impf} in the sense 'to answer', but not in the sense 'to correspond'. In such cases, the set of applicable lemmas is specified directly for the LU (and overrides the set of lemmas specified for the whole lexeme).

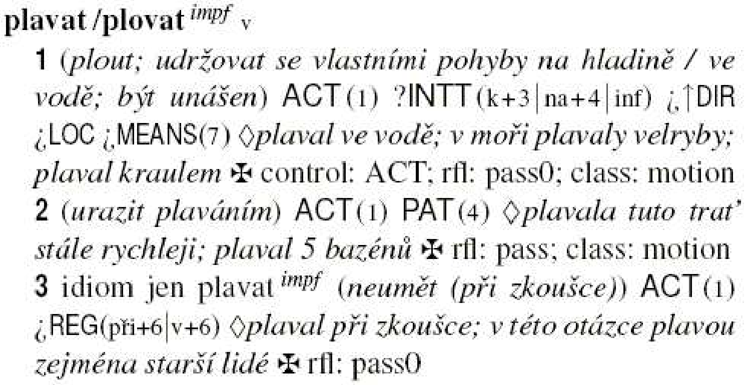
There might be more than one lemma with the same aspect in a lexeme without being lemma variants. Then the aspect flags are distinguished by Arabic numbers, as e.g. in the lexeme *osušovat*^{impf1}, *osoušet*^{impf2}, *osušit*^{pf}, or *odřezávat*^{impf}, *odříznout*^{pf1}, *odřezat*^{pf2} (unique aspect flags are necessary because they serve also for co-indexing the lemmas with example sentences illustrating the usage of the lexeme.

Some verbs (e.g. *informovat*, *charakterizovat*) can be used in different contexts either as imperfective or as perfective. They are called biaspectual verbs.

Within imperfective verbs, there is a subclass of iterative verbs (iter.). Czech iterative verbs are derived more or less in a regular way by affixes such as -*va*- or -*íva*-, and express extended and repetitive actions (e.g., *čítávat*, *chodívat*). In VALLEX 2.0, iterative verbs containing double affix -*va*- (e.g., *chodívávat*) are completely disregarded, whereas the remaining iterative verbs occur as headword lemmas of the relevant lexeme.

2.2.3 Lemma Variants

Lemma variants (many of which are just spelling variants, i.e. orthographic variants) are groups of two or more lemmas that are interchangeable in any context without any change of the meaning (e.g. *dovědět se/dozvědět se*). Usually, the only difference is just a small alternation in the morphological stem, which might be accompanied by a subtle stylistic shift (e.g. *myslet/myslit*, the latter one being bookish). Moreover, although the infinitive forms of the variants differ in spelling, some of their conjugated forms might be identical



plavat /plovat ^{impf} v

1 (*plout; udržovat se vlastními pohyby na hladině / ve vodě; být unášen*) ACT(1) ?INTT(k+3|na+4|inf) ¿↑DIR ¿LOC ¿MEANS(7) ◊*plaval ve vodě; v moři plavaly velryby; plaval kraulem* ✠ control: ACT; rfl: pass0; class: motion
2 (*urazit plaváním*) ACT(1) PAT(4) ◊*plavala tuto trat’ stále rychleji; plaval 5 bazénů* ✠ rfl: pass; class: motion
3 idiom jen plavat ^{impf} (*neumět (při zkoušce)*) ACT(1) ¿REG(při+6|v+6) ◊*plaval při zkoušce; v této otázce plavou zejména starší lidé* ✠ rfl: pass0

Figure 2.4: Sample lexeme with lemma variants.

(*mysli* (imper.sg.) both for *myslet* and *myslit*).

There are rare exceptions when only one of the variants can be used, e.g., *plavat* and *plovat* are usually considered to be variants (see, e.g., [18]). However, in some contexts, only *plavat* can be used (*plaval při zkoušce*, **ploval při zkoušce*). The applicable lemmas must be then listed for the specific LU as in any other case when a LU imposes a further limitation on the set of lexical forms, as shown in Figure 2.4.

2.2.4 Homographs

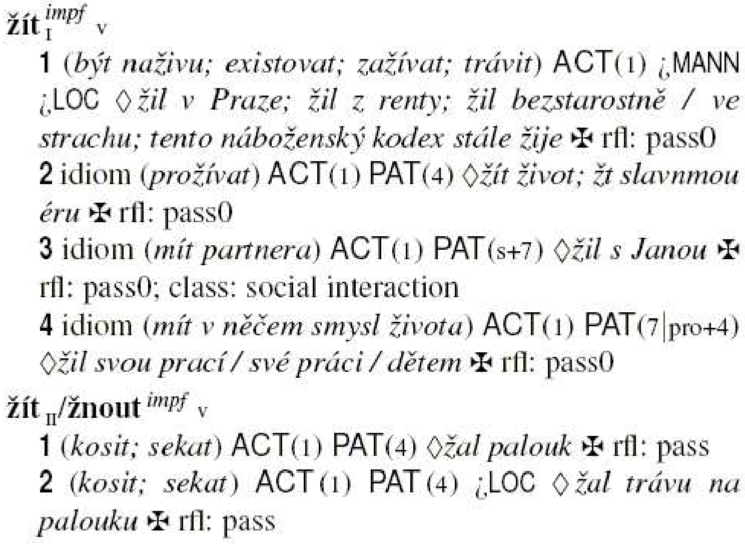
Homographs are lemmas ‘accidentally’ identical in the spelling but considerably different in their meaning (there is no obvious semantic relation between them). They also might differ as to their etymology (e.g. *nakupovat*$_I$ - to buy vs. *nakupovat*$_{II}$ - to heap), aspect (Section 2.2.2) (e.g. *stačit*$_I$ pf. - to be enough vs. *stačit*$_{II}$ impf. - to catch up with), or conjugated forms (*žilo* (past.sg.fem) for *žít*$_I$ - to live vs. *žalo* (past.sg.fem) *žít*$_{II}$ - to mow, see Figure 2.5). In VALLEX 2.0, such lemmas are distinguished by Roman numbering in the subscript. These numbers should be understood as inseparable parts of VALLEX 2.0 lemmas.

2.3 Lexical Units

Each lexeme is formed by a set of lexical units that are assigned to respective lexical forms (represented by their lemmas). Following [1], we understand lexical units (LUs) as “form-meaning complexes with (relatively) stable and discrete semantic properties”. Roughly speaking, LU can be understood as ‘a given word in the given sense’. In the Czech tradition, this concept of LU corresponds to Filipec’s ‘monosemic lexeme’ ([2]).

Within each lexeme in VALLEX 2.0, LUs are numbered by Arabic numbers. In the printed and html versions of the lexicon, the LU entry starts with its number.

The ordering of lexical units is not completely random, but it is not perfectly systematic either. So far, it is based only on the following weak intuition: the primary and/or the most frequent meanings should go first, whereas rare and/or idiomatic meanings should



žít $_{\mathrm{I}}^{impf}$ v

1 (*být naživu; existovat; zažívat; trávit*) ACT(1) ¿MANN ¿LOC ◊*žil v Praze; žil z renty; žil bezstarostně / ve strachu; tento náboženský kodex stále žije* ✠ rfl: pass0

2 idiom (*prožívat*) ACT(1) PAT(4) ◊*žít život; žt slavnmou éru* ✠ rfl: pass0

3 idiom (*mít partnera*) ACT(1) PAT(s+7) ◊*žil s Janou* ✠ rfl: pass0; class: social interaction

4 idiom (*mít v něčem smysl života*) ACT(1) PAT(7|pro+4) ◊*žil svou prací / své práci / dětem* ✠ rfl: pass0

žít $_{\mathrm{II}}$**/žnout** ^{impf} v

1 (*kosit; sekat*) ACT(1) PAT(4) ◊*žal palouk* ✠ rfl: pass

2 (*kosit; sekat*) ACT(1) PAT(4) ¿LOC ◊*žal trávu na palouku* ✠ rfl: pass

Figure 2.5: Sample homograph lexemes.

go last. (We do not guarantee that the ordering of LUs in VALLEX 2.0 exactly matches their frequency in the contemporary language.)

Available information about each LU entry in VALLEX 2.0 is captured by obligatory and optional attributes. The former ones have to be filled with every LU. The latter ones might be empty, either because they are not applicable (e.g. no control can be applicable for verbs without infinitive complementations), or because the annotation was not finished yet (e.g., attribute class; Section 2.5.5).

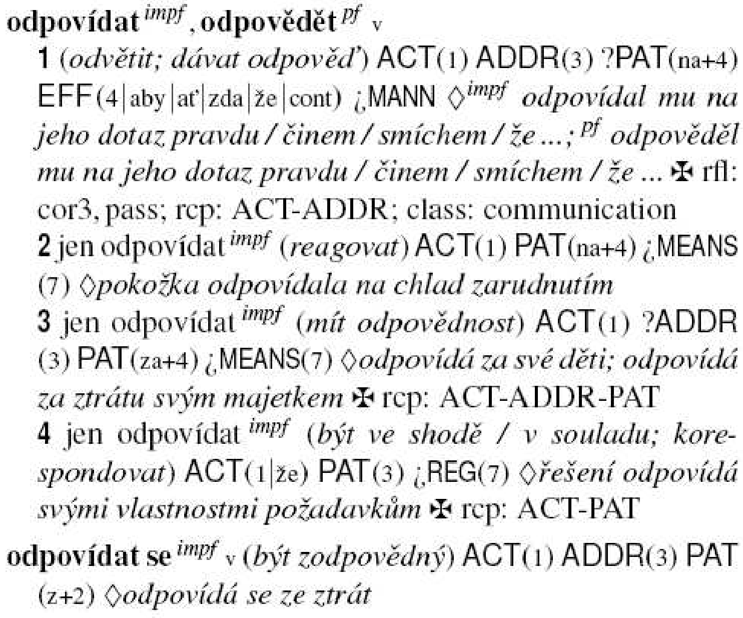
Obligatory LU attributes:

- valency frame (Section 2.4),
- gloss – verb or paraphrase roughly synonymous with the given sense/meaning; this attribute is not supposed to serve as a source of synonyms or even of genuine lexicographic definition – it should be used just as a clue for fast orientation within the word entry!
- example – sentence(s) or sentence fragment(s) containing the given verb used with the given valency frame.

Optional LU attributes:

- flag for idiom (Section 2.5.1),
- information on control (Section 2.5.2),
- possible type(s) of reflexive constructions (Section 2.5.3),
- possible type(s) of reciprocal constructions (Section 2.5.4).
- affiliation to a syntactico-semantic class (Section 2.5.5).

As it can be seen e.g. in Figure 2.6, gloss is located in parentheses at the beginning of every LU entry, and then the valency frame is printed. Example sentence follows the



odpovídat^{impf}, **odpovědět**^{pf} v

1 (*odvětit; dávat odpověď*) ACT(1) ADDR(3) ?PAT(na+4) EFF(4|aby|ať|zda|že|cont) ¿MANN ◊^{impf} *odpovídal mu na jeho dotaz pravdu / činem / smíchem / že ...;* ^{pf} *odpověděl mu na jeho dotaz pravdu / činem / smíchem / že ...* ✠ rfl: cor3, pass; rcp: ACT-ADDR; class: communication

2 jen odpovídat^{impf} (*reagovat*) ACT(1) PAT(na+4) ¿MEANS(7) ◊*pokožka odpovídala na chlad zarudnutím*

3 jen odpovídat^{impf} (*mít odpovědnost*) ACT(1) ?ADDR(3) PAT(za+4) ¿MEANS(7) ◊*odpovídá za své děti; odpovídá za ztrátu svým majetkem* ✠ rcp: ACT-ADDR-PAT

4 jen odpovídat^{impf} (*být ve shodě / v souladu; korespondovat*) ACT(1|že) PAT(3) ¿REG(7) ◊*řešení odpovídá svými vlastnostmi požadavkům* ✠ rcp: ACT-PAT

odpovídat se^{impf} v (*být zodpovědný*) ACT(1) ADDR(3) PAT(z+2) ◊*odpovídá se ze ztrát*

Figure 2.6: Illustration of obligatory and optional attributes within LU entries.

diamond sign, and the optional attributes (if any) are given after the cross sign. If more lemmas are relevant for the given lexeme (as it is often the case because of aspectual pairs), it might be necessary to give more values also in the attribute (especially in the example attribute, see the first LU in the figure). The correspondence between the respective values and the relevant lemmas is captured by superscript labels ^{pf}, ^{impf}, ^{pf1} etc.

2.4 Valency Frames

The core valency information is encoded in the **valency frame**. Within the FGD framework, valency frames (in a narrow sense) consist only of inner participants (both obligatory and optional) and obligatory free modifications. In VALLEX 2.0, valency frames are enriched with quasi-valency complementations. Moreover, a few non-obligatory free modifications occur in valency frames too, since they are typically related to some verbs (or even to whole classes of them) and not to others.[1]

In VALLEX 2.0, a valency frame is modeled as a sequence of frame slots. Each frame slot corresponds to one (either required or specifically permitted) complementation of the given verb.

Note on terminology: in this text, the term 'complementation' (dependent item) is used in its broad sense, not related to the traditional argument/adjunct (complement/modifier) dichotomy.

The following attributes are assigned to each slot:

[1] (The other free modifications can occur with the given verb too, but they are not contained in the valency frame, as their presence in a sentence is not understood as syntactically conditioned in FGD.)

- functor (Section 2.4.1),
- list of possible morphemic forms (realizations) (2.4.2),
- type of complementation (Section 2.4.3).

Some slots tend to occur systematically together. In order to capture this type of regularity, we have introduced the mechanism of slot expansion (Sec. 2.4.4) (full valency frame is obtained after performing these expansions).

2.4.1 Functors

In VALLEX 2.0, functors (labels of ‘deep roles’; similar to theta-roles) are used for expressing types of relations between verbs and their complementations. According to FGD, functors are divided into inner participants (actants) and free modifications (this division roughly corresponds to the argument/adjunct dichotomy). In VALLEX 2.0, we also distinguish an additional group of quasi-valency complementations.

Functors that occur in VALLEX 2.0 are listed in the following tables (for Czech sample sentences see [8], page 43):

Inner participants:

- ACT (actor): *Peter read a letter.*
- ADDR (addressee): *Peter gave Mary a book.*
- PAT (patient): *I saw him.*
- EFF (effect): *We made her the secretary.*
- ORIG (origin): *She made a cake from apples.*

Quasi-valency complementations:

- DIFF (difference): *The value of shares has risen by 100%.*
- OBST(obstacle): *The boy stumbled over a stump.*
- INTT (intent): *He came there to look for Jane.*

Free modifications:

- ACMP (accompaniment): *Mother came with her children.*
- 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.*
- CRIT (criterion): *Peter has to do it exactly according to directions.*
- 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.*
- EXT (extent): *The temperatures reached an all time high.*
- HER (heritage): *He named the new villa after his wife.*
- LOC (locative): *He was born in Italy.*
- MANN (manner): *They did it quickly.*
- MEANS (means): *He wrote it by hand.*
- RCMP (recompense): *She bought a new shirt for 25 $.*
- REG (regard): *With regard to George she asked his teacher for advice.*
- SUBS (substitution): *He went to the theater 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.*
- TTIL (temporal-till-when): *It will last till 5 o'clock.*
- TWHEN (temporal-when): *He will come tomorrow.*

Note 1: Besides the functors listed in the tables above, also value DIR occurs in the VALLEX 2.0 data. It is used only as a special symbol for slot expansion (Sec. 2.4.4).

Note 2: The set of functors as introduced in FGD and used in the Prague Dependency Treebank is richer than that shown above. We do not use its full (current) set in VALLEX 2.0 due to several reasons. Some functors do not occur with verbs at all (e.g., MAT - material, partitive, as *sklenice piva.MAT* - glass of beer), some other functors can occur there but represent other than dependency relations (e.g. coordination, *Jim or.CONJ Jack*). And still others can occur with verbs as well but their behavior is absolutely independent of the head verb; thus they have nothing to do with valency frames (e.g., ATT - attitude, *He did it willingly.ATT*).

2.4.2 Morphemic Forms

In a sentence, each frame slot can be expressed by a limited set of morphemic means which we call forms. In VALLEX 2.0, the set of possible forms is defined either explicitly, or implicitly.

In the first case (explicitly declared forms), the forms are enumerated in a list attached to the given slot (in the case of arguments and quasi-valency complementations, no other forms can be used; in the case of free modifiers, the possible forms are not necessarily limited to those given in the list).

In the second case (implicitly declared forms), no such list is specified because the set of possible forms is implied by the functor of the respective slot (in other words, all forms possibly expressing the given functor may appear).

Explicitly Declared Forms

The list of forms attached to a frame slot may contain values of the following types:

- **Pure (prepositionless) case.** There are seven morphological cases in Czech. In the VALLEX 2.0 notation, we use numbering traditional in the Czech linguistics: 1 - nominative, 2 - genitive, 3 - dative, 4 - accusative, 5 - vocative, 6 - locative, and 7 - instrumental.
- **Prepositional case.** Lemma of the preposition (i.e., preposition without vocalization) and the number of the required morphological case are specified (e.g., *z+2*, *na+4*, *o+6*...). The prepositions occurring in VALLEX 2.0 are the following: *bez, do, jako, k, kolem, mezi, místo, na, nad, o, od, po, pod, podle, pro, proti, před, přes, při, s, u, v, z, za.* ('*jako*' is traditionally considered as a conjunction, but it is included in this list as it requires a particular morphological case in some valency frames).
- **Subordinating conjunction.** Lemma of the conjunction is specified. The following subordinating conjunctions occur in VALLEX 2.0: *aby, ať, až, jak, zda,*[2] *že.*
- **Content clauses.** The abbreviation 'cont' stands for complementations having the form of a content clause (a type of clauses including indirect speech).
- **Infinitive construction.** The abbreviation 'inf' stands for infinitive verbal complementation. 'inf' can appear together with a preposition (e.g. '*než+inf*'), but it happens very rarely in Czech.
- **Construction with adjectives.** Abbreviation 'adj-digit' stands for an adjective complementation in the given case, e.g. adj-1 (*Cítím se slabý* - I feel weak).
- **Constructions with '*být*'** . Infinitive of verb '*být*' (to be) may combine with some of the types above, e.g. *být+adj-1* (e.g. *zdá se to být dostatečné* - it seems to be sufficient).
- **Part of phraseme.** If the set of the possible lexical values of the given complementation is very small (often one-element), we list these values directly (e.g. '*napospas*' for the phraseme '*ponechat napospas*' - to expose).

Implicitly Declared Forms

If no forms are listed explicitly for a frame slot, then the list of possible forms implicitly results from the functor of the slot according to the following (yet incomplete) lists:

- ACMP: bez+2, s+7, společně s+7, spolu s+7, v čele s+7, v souvislosti s+7, ve spojení s+7, včetně+2, ...
- AIM: aby, ať, do+2, k+3, na+4, o+4, pro+4, pro případ+2, proti+3, v zájmu+2, za+4, za+7, že, ...
- BEN: 4, na+4, na účet+2, na úkor+2, na vrub+2, pro+4, proti+3, v+4, ve prospěch+2, v rozporu, s+7, v zájmu+2 ...

[2] Note: form '*zda*' is in fact an abbreviation for the couple of conjunctions '*zda*' and '*jestli*'.

- CAUS: 7, aby, adverb, díky+3, jelikož, ježto, k+7, kvůli+3, na+4, na+6, na základě+2, nad+7, následkem+2, od+2, pod+7, pod náporem+2, pod tíhou+2, pod váhou+2, poněvadž, pro+4, proto, protože, v+6, v důsledku+2, v souvislosti s+7, vinou+2, vlivem+2, vzhledem k+3, z+2, z důvodu+2, za+4, za+7, zásluhou+2, že, …
- CRIT: 7, 2, dle+2, podle+2, na+6, na základě+2, po vzoru+2, přiměřeně+3, v+6, v duchu+2, v rozporu s+7, v souladu s+7, v souhlase s+7, v závislosti na+6, ve shodě s+7, ve smyslu+2, ve světle+2, z titulu+2, …
- DIR1: adverb, od+2, s+2, z+2, ze strany+2, zpod+2, zpoza+2, zpřed+2, …
- DIR2: 7, adverb, kolem+2, cestou+2, mezi+7, napříč+7, po+6, podél+2, přes+4, skrz+4, v+6, …
- DIR3: 2, 7, adverb, do+2, do čela+2, k+3, kolem+2, mezi+4, mimo+4, na+4, na+6, nad+4, naproti+3, okolo+2, po+4, po+6, pod+4, proti+3, před+4, přes+4, směrem do+2, směrem k+3, směrem na+4, v+4, vedle+2, za+4, za+7, …
- EXT: adverb, 2, 4, 7, do+2, kolem+2, k+3, na+4, na+6, nad+4, okolo+2, po+6, pod+7, přes+4, v+4, z+2, za+4, …
- LOC: 2,4, adverb, blízko+2, blízko+3, daleko+2, do+2, kolem+2, mezi+7, mimo+4, na+4, na+6, na úroveň+2, nad+7, naproti+3, nedaleko+2, okolo+2, po+6, po bok+2, poblíž+2, pod+7, podél+2, proti+3, před+7, přes+4, při+6, stranou+2, u+2, uprostřed+2, uvnitř+2, v+6, v čele+2, v oblasti+2, v rámci+2, v řadě+2, vedle+2, za+4, za+7, …
- MANN: 4, 7, adverb, do+2, formou+2, na+4, na+6, nad+4, o+4, po+6, pod+7, proti+3, před+7, při+6, přes+4, s+7, v+4, v+6, v podobě+2, ve formě+2, vedle+2, z+2, za+4, za+7, jak, že …
- MEANS: adverb, 7, cestou+2, díky+3, do+2, na+4, na+6, o+6, po+6, pod+7, pomocí+2, prostřednictvím+2, přes+4, s+7, s pomocí+2, v+6, z+2, za+4, skrz+2, za pomoci+2, že, …
- REG: adverb, 7, bez ohledu na+4, bez zřetele k+3, k+3, kolem+2, na+4, na+6, na téma+2, nad+7, nezávisle na+6, o+6, ohledně+2, po+6, pro+4, před+7, při+6, s+7, se zřetelem k+3, se zřetelem na+4, s ohledem na+4, u+2, v+6, v otázce+2, v případě+2, v rámci+2, v souvislostis+7, ve věci+2, ve vztahu k+3, vůči+3, vzhledem k+3, z+2, z hlediska+2, za+4, …
- SUBS: jménem+2, namísto+2, místo+2, výměnou za+4, za+4,
- TFHL: 4, adverb, do+2, na+4, po+2, pro+4, …
- TFRWH: z+2, od+2, …
- THL: adverb, 2, 4, 7, až, dokud, do+2, na+4, po+4, po dobu+2, přes+4, v+2, za+4, …
- TOWH: adverb, do+2, k+3, na+4, pro+4, …
- TSIN: adverb, od+2, počínaje+7, z+2, …
- TTILL: adverb, do+2, dokud, k+3, než, po+4, …
- TWHEN: 2, 4, 7, adverb, až, do+2, jakmile, k+3, když, kolem+2, koncem+2, mezi+7, na+4, na+6, na závěr+2, než, o+6, okolo+2, po+6, počátkem+2, postupem+2, poté co, před+7, předtím než, při+6, s+7, u příležitosti+2, v+4, v+6, v době+2, v období+2, v průběhu+2, v závěru, z+2, za+2, za+4, začátkem, …

2.4.3 Types of Complementations

Within the FGD framework, valency frames (in a narrow sense) consist only of inner participants (both obligatory[3] and optional) and obligatory free modifications; the dialogue test was introduced by Panevová [11] as a criterion for obligatoriness (see [16]). In VALLEX 2.0, valency frames are enriched with quasi-valency complementations. Moreover, a few non-obligatory free modifications occur in valency frames too, since they are typically related to some verbs (or even to whole classes of them) and not to others.

The attribute 'type' is attached to each frame slot and can have one of the following values: 'obl' or 'opt' for inner participants and quasi-valency complementations, and 'obl' or 'typ' for free modifications. In the printed version, optional complementations are marked with '?', whereas typical complementations are marked with '¿'.

2.4.4 Slot Expansion

Some slots tend to occur systematically together. For instance, verbs of motion can be often modified with direction-to and/or direction-through and/or direction-from modifier. We decided to capture this type of regularity by introducing the abbreviation flag for a slot. If this flag is set (in the VALLEX 2.0 notation it is marked with an upward arrow), the full valency frame is obtained after slot expansion.

If one of the frame slots is marked with the upward arrow (in the XML data, attribute 'abbrev' is set to 1), then the full valency frame will be obtained after substituting this slot with a sequence of slots as follows:

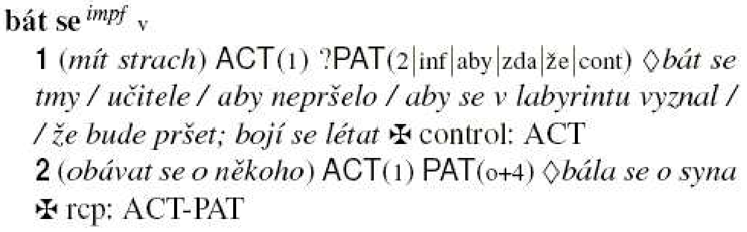
- $\uparrow\mathrm{DIR}^{typ} \rightarrow \mathrm{DIR1}^{typ}\ \mathrm{DIR2}^{typ}\ \mathrm{DIR3}^{typ}$
- $\uparrow\mathrm{DIR1}^{obl} \rightarrow \mathrm{DIR1}^{obl}\ \mathrm{DIR2}^{typ}\ \mathrm{DIR3}^{typ}$
- $\uparrow\mathrm{DIR2}^{obl} \rightarrow \mathrm{DIR1}^{typ}\ \mathrm{DIR2}^{obl}\ \mathrm{DIR3}^{typ}$
- $\uparrow\mathrm{DIR3}^{obl} \rightarrow \mathrm{DIR1}^{typ}\ \mathrm{DIR2}^{typ}\ \mathrm{DIR3}^{obl}$
- $\uparrow\mathrm{THL}^{typ} \rightarrow \mathrm{TSIN}^{typ}\ \mathrm{THL}^{typ}\ \mathrm{TTIL}^{typ}$

2.5 Optional LU Attributes

2.5.1 Idioms

When building VALLEX, we have focused mainly on primary or usual meanings of verbs. We also noted many LUs corresponding to peripheral usages of verbs. However their coverage in VALLEX might not be complete. We call such LUs idiomatic and mark them with the label 'idiom' (e.g. LUs 4 and 5 in Figure 2.1). An idiomatic frame is tentatively characterized either by a substantial shift in meaning (with respect to the primary sense), or by

[3] It should be emphasized that in this context the term obligatoriness is related to the presence of the given complementation in the deep (tectogrammatical) structure, and not to its (surface) deletability in a sentence (moreover, the relation between deep obligatoriness and surface deletability is not at all straightforward in Czech).



bát se ^{impf} v

1 (*mít strach*) ACT(1) ?PAT(2|inf|aby|zda|že|cont) ◇*bát se tmy / učitele / aby nepršelo / aby se v labyrintu vyznal / / že bude pršet; bojí se létat* ✠ control: ACT

2 (*obávat se o někoho*) ACT(1) PAT(o+4) ◇*bála se o syna* ✠ rcp: ACT-PAT

Figure 2.7: Sample lexeme with control in its first LU.

a small and strictly limited set of possible lexical values in one of its complementations, or by occurrence of another type of irregularity or anomaly.

2.5.2 Control

The term ‘control’ relates in this context to a certain type of predicates (verbs of control) and two coreferential expressions, a ‘controller’ and a ‘controllee’. In VALLEX 2.0, control is captured in the data only in the situation in which a verb has an infinitive modifier (regardless of its functor). Then the controllee is an element that would be a ‘subject’ of the infinitive (which is structurally excluded on the surface), and controller is the co-indexed expression. In VALLEX 2.0, the type of control is stored in the frame attribute ‘control’ as follows:

- if there is a coreferential relation between the (unexpressed) subject (‘controllee’) of the infinitive verb and one of the frame slots of the head verb, then the attribute is filled with the functor of this slot (‘controller’), see Figure 2.7;
- otherwise (i.e., if there is no such co-reference), value ‘ex’ is used.

Examples:

- *pokusit se* (to try) - control: ACT,
- *slyšet* (to hear), e.g. *slyšet někoho přicházet* (to hear somebody coming) - control: PAT,
- *jít*, in the sense *jde to udělat* (it is possible to do it) - control: ex.

2.5.3 Reflexivity

The optional attribute reflexivity (abbreviation ‘rfl’) indicates possible syntactic functions of the reflexive particles/pronouns *se* or *si*.

The reflexive particles/pronouns *se* or *si* are used in Czech as formal means expressing the following syntactic constructions:

- derived diatheses: the particle *se* is a part of the reflexive passive verb form:
  - for transitive verbs (e.g *plány se připravují* - plans are prepared); marked with the label ‘pass’,

  - for intransitive verbs (e.g. *pátrá se po zloději* - a thief is being looked for; *v neděli se chodí do kostela* - on Sundays one visits the church); marked with the label ‘pass0’.

- grammatical coreference: the pronouns *se* or *si* stands for an inner participant that is coreferential with Actor (e.g. *mýt se* (to wash oneselfs) - coreference between ACT and PAT (in Accusative); *podřídit si zaměstnance* (to bring under the employees) - coreference between ACT and ADDR in dative); marked with the labels ‘cor3’ (in the case of *si*) or ‘cor4’ (in the case of *se*)

Note that the attribute reflexivity does not cover reflexive verb forms where reflexive particles *se* or *si* are parts of the infinitive forms, i.e. true reflexive (e.g. *bát se* (to fear), *smát se* (to laugh)) as well as derived reflexive (e.g. *odpovídat se* (to account), *šířit se* (to spread), *vrátit se* (to return)) (as already discussed in Section 2.2.1), nor the reciprocical function of *se* or *si* pronouns (see the following Section).

2.5.4 Reciprocity

Reciprocity is understood as a possibility of (two or more) valency complementations to be in relations with each other that may be viewed symmetrically (and their roles are interchangeable).

In Czech, if Actor and some other complementation are reciprocal, then the reflexive verb form is used and these two complementations are expressed either as a coordinated nominal group (as in *Petr a Marie se hádali* - Peter and Mary argued (with one another))), or as a plural noun (*Přátelé se navštěvují* - Friends visit each other), possibly with additional adverbs *spolu*, *navzájem*, . . .

If Actor is not affected, the reciprocity may follow from the plural form or coordination (with no other formal sign), as in *Seznámil je* - he introduced them (to each other).

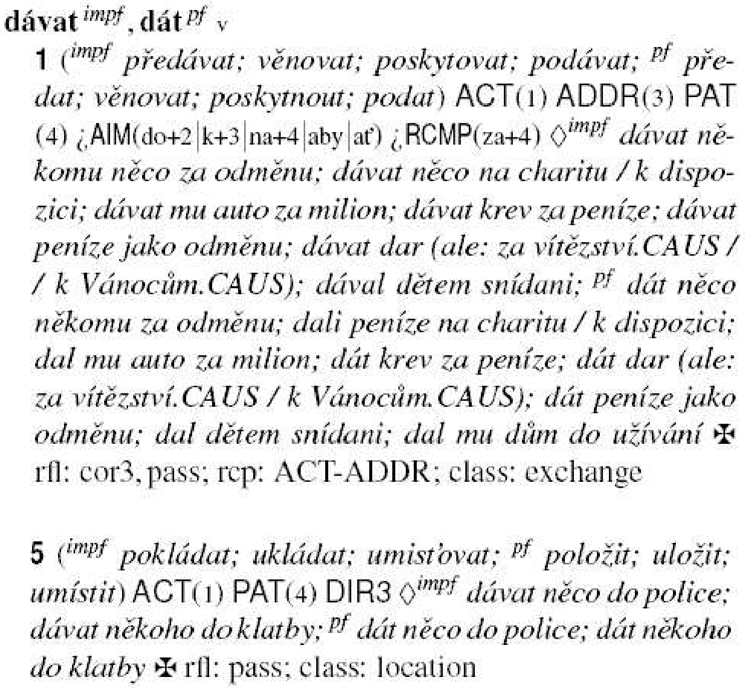
The possibility of reciprocal usage is indicated in the attribute reciprocity (‘rcp’ for short), the value of which is a pair (or triple) of functors involved, e.g. ACT-ADDR for *hádat se* (to argue) (*neustále se spolu hádali* - they argued with each other all the time), or ACT-ADDR-PAT for *mluvit* (to talk) (*mluví spolu o sobě* - they talked with each other about themselves).

In the case of derived reflexive lexemes of inherently reciprocal verbs (with the obligatory complementation with the form s+7), both LUs for irreflexive and reflexive lexemes have assigned attribute ‘rcp’. Example:

- ACT-PAT for *navštěvovat, navštívit* (impf: *navštěvovali se vzájemně*, pf: *navštívit se navzájem* - they visited each other),
- ACT-PAT for *navštěvovat se* (*navštěvovali se pravidelně celá léta* - they visited each other for all the years).

2.5.5 Class

Some frames are assigned semantic classes like ‘motion’, ‘exchange’, ‘communication’, ‘perception’, etc. However, we admit that this classification is tentative and should be



dávat ^{impf}, **dát** ^{pf} v

1 (^{impf} *předávat; věnovat; poskytovat; podávat;* ^{pf} *předat; věnovat; poskytnout; podat*) ACT(1) ADDR(3) PAT(4) ¿AIM(do+2|k+3|na+4|aby|ať) ¿RCMP(za+4) ◊^{impf} *dávat někomu něco za odměnu; dávat něco na charitu / k dispozici; dávat mu auto za milion; dávat krev za peníze; dávat peníze jako odměnu; dávat dar (ale: za vítězství.CAUS / / k Vánocům.CAUS); dával dětem snídani;* ^{pf} *dát něco někomu za odměnu; dali peníze na charitu / k dispozici; dal mu auto za milion; dát krev za peníze; dát dar (ale: za vítězství.CAUS / k Vánocům.CAUS); dát peníze jako odměnu; dal dětem snídani; dal mu dům do užívání* ✠ rfl: cor3, pass; rcp: ACT-ADDR; class: exchange

5 (^{impf} *pokládat; ukládat; umisťovat;* ^{pf} *položit; uložit; umístit*) ACT(1) PAT(4) DIR3 ◊^{impf} *dávat něco do police; dávat někoho do klatby;* ^{pf} *dát něco do police; dát někoho do klatby* ✠ rfl: pass; class: location

Figure 2.8: Illustration of the attribute class.

understood merely as an intuitive grouping of frames, rather than a properly defined ontology. The motivation for introducing such semantic classification in VALLEX 2.0 was the fact that it simplifies systematic checking of consistency and allows for making more general observations about the data.

Bibliography

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

[2] Josef Filipec and František Čermák. *Česká lexikologie*. Academia, Praha, 1985.

[3] Eva Hajičová, Barbara H. Partee, and Petr Sgall. *Topic-Focus Articulation, Tripartite Structures, and Semantic Content*, volume 71 of *Studies in Linguistics and Philosophy*. Kluwer, Dordrecht, 1988.

[4] Markéta Lopatková. Valency in the Prague Dependency Treebank: Building the Valency Lexicon. *The Prague Bulletin of Mathematical Linguistics*, (79–80):37–60, 2003.

[5] Markéta Lopatková and Jarmila Panevová. Recent developments in the theory of valency in the light of the Prague Dependency Treebank. In Mária Šimková, editor, *Insight into Slovak and Czech Corpus Linguistic*, pages 83–92. Veda Bratislava, Slovakia, 2005.

[6] Markéta Lopatková, Zdeněk Žabokrtský, and Karolína Skwarska. Valency Lexicon of Czech Verbs: Alternation-Based Model. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, volume 3, pages 1728–1733. ELRA, 2002.

[7] Markéta Lopatková, Zdeněk Žabokrtský, Karolina Skwarska, and Václava Benešová. VALLEX 1.0 Valency Lexicon of Czech Verbs. Technical Report TR-2003-18, UFAL/CKL MFF UK, Prague, 2003.

[8] Markéta Lopatková, Zdeněk Žabokrtský, Karolína Skwarska, and Václa va Benešová. Tektogramaticky anotovaný valenční slovník českých sloves. Technical Report TR-2002-15, ÚFAL/CKL MFF UK, Prague, 2002.

[9] Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, and Lucie Kučová. Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka. Technical Report TR-2005-28, ÚFAL MFF UK, Prague, Prague, 2005.

[10] Karel Pala and Pavel Ševeček. Valence českých sloves. In *Sborník prací FFBU*, pages 41–54, Brno, 1997.

[11] Jarmila Panevová. On Verbal Frames in Functional Generative Description. *The Prague Bulletin of Mathematical Linguistics*, (22):3–40, 1974.

[12] Jarmila Panevová. *Formy a funkce ve stavbě české věty*. Academia, Praha, 1980.

[13] Jarmila Panevová. More Remarks on Control. *Prague Linguistic Circle Papers, John Benjamins*, 2:101–120, 1996.

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

[15] Jarmila Panevová. Znovu o reciprocitě. *Slovo a slovesnost*, 68, 2007.

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

[17] Hana Skoumalová. *Czech syntactic lexicon*. PhD thesis, Univerzita Karlova, Filozofická fakulta, 2001.

[18] *Slovník spisovného jazyka českého*. Praha, 1964.

[19] *Slovník spisovné češtiny pro školu a veřejnost*. Academia, Praha, 1978.

[20] Naďa Svozilová, Hana Prouzová, and Anna Jirsová. *Slovesa pro praxi*. Academia, Praha, 1997.

[21] Naďa Svozilová, Hana Prouzová, and Anna Jirsová. *Slovník slovesných, substantivních a adjektivních vazeb a spojení*. Academia, Praha, 2005.

[22] Zdeněk Žabokrtský. *Valency Lexicon of Czech Verbs (PhD thesis)*. PhD thesis, Charles University, Prague, Czech Rep., 2005.

Part II

LEXEME ENTRIES