Building the PDT-Vallex valency lexicon

Zdeňka Urešová
Charles University in Prague
Institute of Formal and Applied Linguistics
Czech Republic
uresova@ufal.mff.cuni.cz

Abstract

In our contribution, we relate the development of a richly annotated corpus and a computational valency lexicon. Our valency lexicon, called PDT-Vallex (Hajič et al., 2003) has been created as a "byproduct" of the annotation of the Prague Dependency Treebank (PDT) but it became an important resource for further linguistic research as well as for computational processing of the Czech language.

We will present a description of the verbal part of this lexicon (more than 5300 verbs with 8200 valency frames) that has been built on the basis of the PDT corpus. Rigorous approach and the linking of each verb occurrence to the valency lexicon has made it possible to verify and refine the very notion of valency as introduced in the Functional Generative Description theory (Sgall et al., 1986; Panevová, 1974-5).

Every occurrence of a verb in the corpus contains a reference to its valency frame (i.e., to an entry in the PDT-Vallex valency lexicon). The annotators insert the verbs (verb senses) found in the course of the annotation and their associated valency frames into the lexicon, adding an example (or more examples) of its usage (directly from the corpus). They also insert a note that refers to another verb that has one of its valency frames related to the current one (a synonym/antonym, an aspectual counterpart, etc.).

A functor as well as its surface realization is recorded in every slot of each valency frame. The mapping between the valency frame and its surface realization is generally quite complex (Hajič and Urešová, 2003). The surface realizations through the morphemic case, preposition and a case, and subordinate sentence are the most common.

The valency frame is fully formalized to allow for automatic computerized processing of the valency dictionary entries. Verb complementations are marked for obligatoriness, and their surface realization is attached. The realization of inner participants (arguments) is always given in full, since there is no "standard" or "default" realization; free modifications' (adjuncts') realization need not be specified.

The PDT-Vallex is available as part of the PDT version 2 published by the Linguistic Data Consortium (http://www.ldc.upenn.edu, LDC2006T01).[5]

1 Introduction

Before we will concentrate on the PDT-Vallex, let us make a little digression into the background of this lexicon. There are two main sources for building the PDT-Vallex. Firstly, it is a corpus, in our case the Prague Dependency Treebank (PDT). Secondly, it is a well developed valency theory; we work with the Functional Generative Description valency theory (FGDVT).

1.1 The Prague Dependency Treebank (PDT)

The computational valency lexicon PDT-Vallex has been built on the basis of the Prague Dependency Treebank (PDT). The PDT is a long term open ended project for manual

annotation of Czech texts. This project started in 1996 in the Institute of Formal and Applied Linguistics and in the Centre for computational linguistics at the Faculty of Mathematics and Physics, Charles University, Prague. The project has two main purposes: the first important goal is to test and preserve the linguistic theory which is behind the whole project and the second is to apply and test machine learning methods for developing POS and morphological taggers, dependency parsers, NLG tools, programs for coreference resolution and many other.

Our source of raw texts in electronic form was the Institute of the Czech National Corpus.[1] The data in the PDT are non-abbreviated articles from Czech newspapers and journals. The PDT is based on the framework of the Functional Generative Description (FGD), which has been developed within the Prague School of Linguistics by P. Sgall and his collaborators (Sgall et al., 1986). The FGD is dependency-oriented with a stratificational (or *layered*) approach to a systematic description of a language. The PDT annotation is very rich in linguistic information; the corpus itself now uses the latest annotation technology.[2]

The same text is annotated at different but linked grammatical levels (see Fig. 1). It starts with morphology (morphological layer), continues with surface syntax (analytical layer) and goes to a combination of "deep" syntax and semantics (tectogrammatical layer).

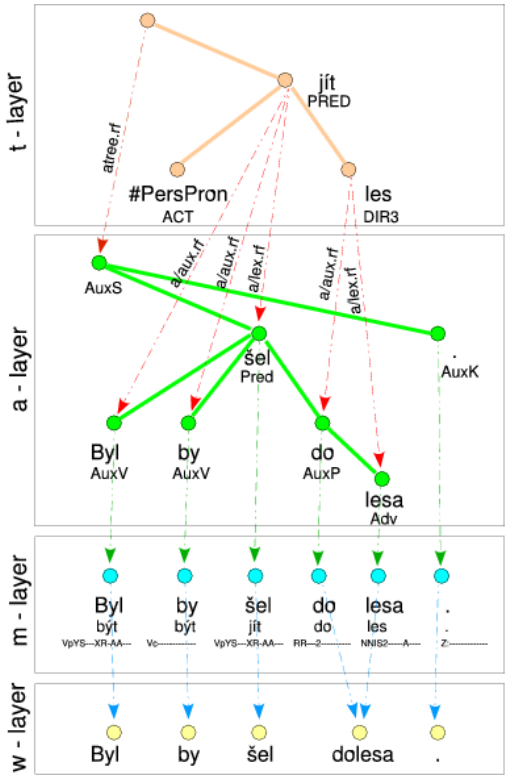

Figure 1: Linking the layers

On the lowest, morphological layer (with 2 million tokens annotated), several attributes, the most important of which are morphological lemma and tag, are being annotated. At the analytical layer (with about 1.5 million tokens annotated), a sentence is represented as a rooted ordered tree with labelled nodes and edges – i.e., a dependency tree. Nodes of this tree represent tokens as they are found in the original sentence. No node is added or deleted. Edges usually represent relation of formal dependency. In addition, an analytical function capturing the type of dependency relation between the child and its parent is added (Hajič et al., 1999). The last (so far) annotated layer in the PDT is the tectogrammatical layer (Mikulová et al., 2005). The tectogrammatical layer (the layer of "deep" underlying syntax, or the layer of "linguistic

meaning") is in the focus of the FGD theory. It represents the deep (underlying) structure of a sentence. The tectogrammatical representation captures tectogrammatical structure and functors, topic–focus articulation and coreference. At this layer, nodes represent only non-function ("autosemantic") words. They carry several attributes; one of the most important ones is the functor capturing the tectogrammatical dependency relation between a dependent and its governor. The functor is strongly related to the notion of valency. Purely morphological and syntactic attributes, such as morphological lemma, case, number, gender, syntactic function (subject, object, …) and many others are not present, but they can be easily retrieved by following the inter-layer links depicted in Fig. 1 by the dashed arrows.

1.2 The Functional Generative Description Valency Theory (FGDVT)

The valency concept in the PDT-Vallex stems from the main principles of the "standard"[3] FGD valency theory which is a substantial part of the FGD. The FGDVT has been being developed since the 70s, concentrating firstly especially on verbs and extended later to other parts of speech. The theoretical description of FGDVT is summarized mainly by Panevová (Panevová, 1974-75, 1998, 1999, 2002).

This theory is dependency oriented and it "operates" on the tectogrammatical layer. It combines the syntactic and semantic approach for distinguishing valency elements. The verb is considered to be the core of the sentence (or clause, as the case may be). The relation between the dependent and its governor at the tectogrammatical layer is labelled by a *functor*. For a full list of all dependency relations and their labels, i.e., the functors, as they are used in the PDT (based on those described and used in the FGDVT) see (Mikulová et al., 2005).

FGDVT works with a systematic classification of verbal valency modifications along two axes. The first axis concerns the opposition between *inner participants* (arguments) and *free modifications* (adjuncts). The other axis relates to the distinction between *obligatory* and *optional* complementations.

There are five "inner participants" (arguments): Actor/Bearer (ACT), Patient (PAT), Addressee (ADDR), Origin (ORIG) and Effect (EFF). Which functors are arguments have been determined according to two criteria. The first one says that arguments can occur at most once as a dependent at a single occurrence of a particular verb (excluding apposition and coordination). According to the second criterion, each of them can modify only a relatively closed class of verbs.

Out of the five argument types, FGDVT states that the first two are connected with no specific semantics, contrary to the remaining three ones. The first argument is always the Actor (ACT), the second one is always the Patient (PAT). The Addressee (ADDR) is the semantic counterpart of an indirect object that serves as the recipient or simply the "addressee" of the action described by the verb. Effect (EFF) is the semantic counterpart of the second indirect object describing typically the result of the action (or the contents of an indirect speech, for example, or a state as described by a verbal attribute – the complement). Origin (ORIG) also comes from the second (or third or fourth) indirect object, describing, not surprisingly, the origin of the action (in the "creation" sense, such as *to build from metal sheets*, not in the directional sense).

FGDVT has further adopted the concept of shifting of „cognitive roles". According to this special rule, semantic Effect, semantic Addressee and/or semantic Origin are being shifted to the Patient position in case the verb has only two arguments. In the sentence *Peter has dug a hole,* the semantic Effect (*a hole*) happens to be a Patient; similarly, in the sentence *The teacher asked the pupil* the semantic Addressee (*the pupil*) is shifted to the Patient position. Similarly, in *The book came out* the deep object (Patient, *the book*) is shifted to the Actor position if there is no apparent Actor present. This rule, when viewed from the annotation point of view, helps to keep consistency at the expense of lower "semantic precision".

More examples of the use of the shifting rule (English-only examples given):

(1) PAT shifted to the position of ACT: *The new law.ACT came into force*

(2) ADDR shifted to the position of PAT: *she liked him.PAT*

(3) ADDR shifted to the position of PAT, EFF stays in its slot: *elect her.PAT as a president.EFF*

(4) EFF shifted to the position of PAT: *to build a group.PAT*

(5) ORIG shifted to the position of PAT: *the sunflower grows out of seed.PAT*.

There are only some specific cases where shifting does not apply.

As for their morphemic realization, arguments are governed and they can be both obligatory and optional.

Examples of arguments:

(6) Marta.ACT zvětšila svou zahradu.PAT z 20.ORIG na 40.EFF hektarů.

(7) Martha.ACT enlarged her garden.PAT from 20.ORIG to 40.EFF hectares.

(8) Jana.ACT poslala Karlovi.ADDR dopis.PAT

(9) Jane.ACT sent Charles.ADDR a letter.PAT

(10) Jirka.ACT rozřezal jablko.PAT na kousky.EFF.

(11) George.ACT cut the apple.PAT into pieces.EFF.

The repertory of adjuncts (free modifications) is much larger than that of arguments. FGD distinguishes about 50 types of adjuntcs (for the full list of adjuncts see Mikulová et al., 2005). Adjuncts are always determined semantically; their set might be divided into several subclasses, such a temporal (TWHEN, TSIN, TTILL, TFL, TFHL, THO, TPAR, TFRWH, TOWH), local (LOC, DIR1, DIR2, DIR3), causal (such as CAUS for cause, AIM, CRIT for 'according to', etc.) and other free modifications (MANN for general 'manner', ACMP for accompaniment, EXT for extent, MEANS, INTF for intensifier, BEN for benefactor, etc.) Adjuncts may be seen as deep-layer counterparts of surface adverbial complementations. An adjunct of the same type can occur more than once with a particular occurrence of the verb and adjuncts can modify in principle any verb – this is also where their name ('free modifications') comes from. Unlike arguments, morphemic realization of adjuncts is rarely if ever restricted by the particular verb.

Examples of adjuncts:

(12) Michal se včera.TWHEN ve škole.LOC choval divně.MANN.

(13) Michael behaved yesterday.TWHEN in school.LOC funny.MANN.

(14) Anna jela po Praze.DIR2 vlakem.MEANS na Filozofickou fakultu.DIR3, aby dostala.AIM studentskou kartu.

(15) Anna went through Prague.DIR2 by the train.MEANS to the Faculty of Arts.DIR3 to get a student card.AIM.

Due to this "free nature" of adjuncts, only the presence of arguments (obligatory or optional) and obligatory adjuncts is considered necessary at any verbal valency frame (FGDVT is thus said to use the notion of valency in its "narrow" sense): optional adjuncts are not listed in the valency frame. As mentioned above, both arguments and adjuncts can be in their relation to a particular word either obligatory (that means obligatorily present at the tectogrammatical level) or optional (that means not necessarily present in each sentence where the verb is used). It must be said that this definition of obligatoriness and optionality does not cover surface deletions – they can appear almost anywhere – but only *semantically* necessary elements.

Since the surface appearance of a complementation does not really help to distinguish between obligatory and optional elements, other criteria must be used. Specifically, the '*dialogue test*' is used. It is a method based on a question about something that is supposed to be known to the speaker because it follows from the meaning of the verb: if the speaker can answer hearer's follow-up wh-question about a given complementation "I don't know" without confusing the hearer, it means that the given modification is semantically optional. On the other hand, if the answer "I don't know" is not disruptive in the (assumed) conversation, then the given modification is considered to be semantically obligatory.

For example (for the verb '*to leave*'), in '*John left*' the speaker must know '*from where*' (otherwise, he or she would highly probably use another verb, such as '*to go*', '*to travel*', etc.). Thus '*from where*' is an obligatory modification. However, in the same situation, the speaker does not really have to know '*to where'* John left (if the destination were the core of his communication, he would use a verb such as '*to arrive*'). In this case, the dialogue test concludes that 'to *where'* modification is optional. (For detailed information about this matter see Urešová, 2005.)

2 Building the PDT-Vallex

The valency issue forms the core ingredient in the annotation of the PDT. The knowledge of valency frames plays the most important role during the process of the annotation of the tectogrammatical layer. As it has been already mentioned, each dependency relation at this layer is labelled by a functor, and this label is, for all verb (and some noun) dependents directly determined by the verb (noun) sense and its valency. It was thus clear that valency would play a crucial role here. Of course, it would be ideal to have the complete valency lexicon at our disposal before starting the annotation process, but this was not the case.

For the purpose of annotation of the tectogrammatical layer it was thus necessary to build the PDT-Vallex essentially from scratch, notwithstanding some small sets of verbs with initial attempts at rather informal valency description attached to them. The annotators simply had to add entries as they went on with the structural annotation. In fact, they were trained to add entries essentially *only* when they encountered new verb meaning in the actual data. This "bottom up", practical approach to the forming of the valency lexicon enabled for the first time the confrontation of the already developed valency theory (FGDVT) and real usage of language (Czech written texts). Instead of "creating" examples for each sense and valency frame based on pure intuition or using limited excerpts, the lexicon drew upon the real texts, upon real corpus data. This not only resulted in examples taken from real language usage (albeit often simplified), but mainly, it has naturally opened new questions and problems when confronted with the FDGVT. Answers to these questions have often lead to more precise boundaries between groups of analogical cases, but sometimes also to deeper questions about all issues related to valency. These issues had to be solved also on the go, while having mainly

consistency and effectiveness of annotation in mind. For more information about problems with solving valency-related problems in the PDT see (Urešová, 2005).

Making the PDT-Vallex updated (always as soon as questions have been answered and problems solved) served well for keeping inter-annotator consistency high during the process of corpus annotation. After the tectogrammatical annotation process has ended, we have revised the whole PDT-Vallex for cross-entry consistency (for groups of verbs having similar meanings, for aspectual counterparts etc.), and then the corpus has been amended and corrected using the final version of the lexicon.

The lexicon served also for rigorous, automatic cross-checking of the annotated PDT data against this newly built lexicon in the final stages of quality control of the data before their public release.

3 The contents of the PDT-Vallex

Only those words (verbs, nouns, adjectives and adverbs) and their senses which occurred in the annotated data are recorded in the PDT-Vallex. The lexicon now contains over 12000 different words: almost 7500 verbs, 3800 nouns, 800 adjectives and a few adverbs. The total number of valency frames is almost 16000.

The following valency frames of different parts of speech found in the PDT data are included in the PDT-Vallex:

1. Valency frames of all semantic verbs.
2. Valency frames of semantic nouns which are constituents of complex predicates.
3. Valency frames of semantic nouns, adjectives and adverbs that have at least one daughter node labelled by one of the following functors: ACT, PAT, ADDR, EFF or ORIG.
4. Valency frames for non-verbal idioms containing as the governing node either a semantic adverb or a semantic noun.
5. Valency frames of non-verbal idioms containing as the governing node a semantic verbal noun (cf. Mikulová et al., 2005)

In this paper, we concentrate on the main, i.e. the verbal, part of the lexicon. The PDT-Vallex contains almost 7500 verbs (out of which 5510 have been used in the PDT in ca. 8 500 meanings). Every verb occurrence in the PDT has a link to one of the valency frame in the PDT-Vallex lexicon. For example for the verb *uzavřít* (*to close*), there are two valency frames in the lexicon. In Figure 2 the verb *uzavřít* (*to close*) is used in three different sentences in the PDT. The first two occurrences are linked with the second valency frame with the basic meaning of *to_close* (which has the usual transitive frame with two arguments, ACT and PAT) and the third occurrence of *to close* is linked with the first valency frame (which represents the light-verb meaning – denoted here with the CPHR functor in its frame). The arguments -- Actor and a Patient in the first two sentences and a CPHR functor in the third sentence – are then implicitly linked to the lexicon as well.

The link is stored in an attribute called `val-frame.rf`. This attribute keeps the valency frame identifier which was used at the labelled node (as attested by its dependent nodes which are the complements corresponding to the slots stored in the valency frame). The t-lema used in the data and t-lema kept in the lexicon match as well, of course.

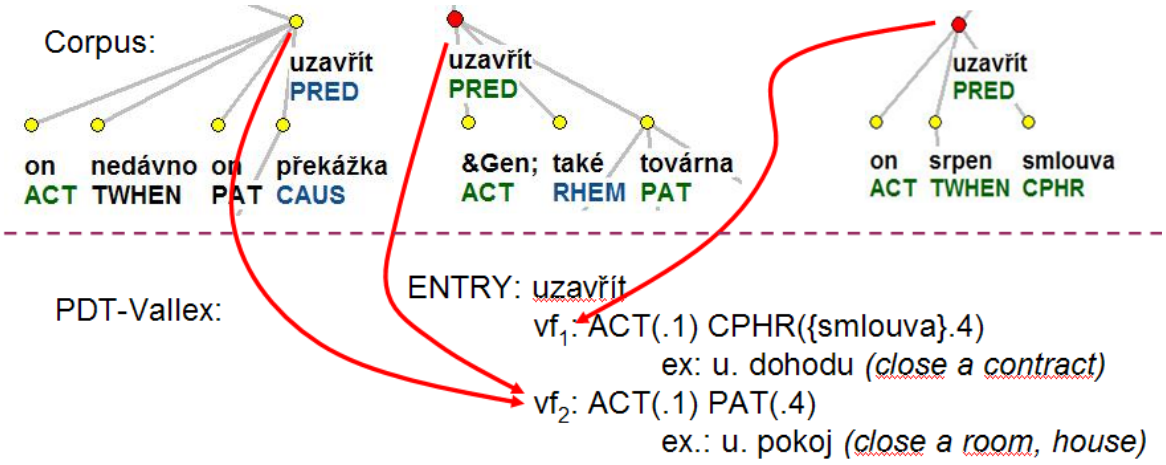

Figure 2. Links between the corpus and the PDT-Vallex entries.

3.1 The valency entry in the PDT-Vallex

The valency entry in the lexicon contains `t-lema` (which is the "headword", by which the valency frames are grouped, indexed, and sorted) and its valency frame.

A valency frame corresponds to one verb sense. The term '*(verb) sense*' is used rather intuitively here; we have worked with the notion of concrete, abstract and idiomatic (phraseological) sense, but this distinction is not explicitly marked in the PDT. It has served rather for the development of consistent frames across synonymous or nearly synonymous verbs (again, without really defining the exact nature of synonymy). The verb senses are further distinguished, but in most cases we have ended at quite coarse-grained set of senses, typically differentiated also by their complementation structure. Only rarely, and only in the clearest cases of two or more senses being distinct, we have 'split' an entry into two or more valency frames with identical complement structure.

The following specification is found in the valency frame :

(a) The number of valency frame members – the number of valency frame members is fixed (zero, i.e. no complementation is also possible, such as for one sense of *pršet* '*to rain*').

(b) The labels of valency frame members ('functors') – the valency frame members are distinguished by them..

(c) The obligatoriness feature of the valency frame members – in accordance with the FGDVT, members of the valency frame are designed as either obligatory or optional.

(d) The surface form of valency frame members – the basic (or '*canonical*') valency frame describing the surface form of verb complementations when the verb is used in the active voice is recorded. The canonical valency frame thus stands for the *primary diathesis*[4], however, frames for secondary diatheses can be created automatically with specific transformation rules we have developed (Pajas and Urešová, 2009).

(e) Examples – any concrete lexical realization of the particular valency frame is exemplified through an appropriate example which comprises an understandable fragment of a Czech sentence originating almost exclusively from the PDT corpus. The example in question might be slightly adjusted for transparency purposes in some cases, or it might even be made up (very rarely, though). If there is any doubt which word from the given example refers to which valency frame member, such a word is

informally labelled in the text of the example by the appropriate functor as a visible hint.

(f) Notes – notes help the annotators or users of the dictionary distinguish the individual valency frames inside the valency entry based on their sense. Typically, synonyms, antonyms and aspectual counterparts serve as notes. Such notes are not considered – unlike the elements described above in (a) – (e) compulsory; however, they are listed almost in every case except in the valency frames with an idiomatic (phraseological) meaning where the sense is already clear from the valency realization.

To summarize, the lexicon user (and computer programs working with it) learns from every valency frame as stored in the PDT-Vallex the following information: how many members the frame has, how the individual members (complementations) are called (their label – 'functor') and what are the additional features of the frame members (obligatory/optional, sense description, surface form).

Unlike in some traditional Czech valency theories (Daneš 1985, Pauliny 1943), PDT-Vallex has no notion of left- or right valency. No explicit markup of the position of the valency complementations relative to the position of the verb is given – after all, Czech is a free word order language (more precisely, word order is not determined by syntax but by the information structure of the sentence).

3.2 The valency frame and meaning in the PDT-Vallex

The valency frame stores valency complementations (valency frame members) of the given verb. One verb has usually more meanings and therefore more separate valency frames – one valency frame relates to one verb meaning. Every verb has as many valency frames as it has meanings.

In the examples that follow, we use the full valency frame specification as recorded in the PDT-Vallex lexicon; for the description of the functors, see Chapter 3.4, and for the description of the surface form realization, see Chapter 3.6.

(16) přišít: [stitch]
přišít¹ [sew] – *přišít knoflík na košili* [sew a button on the shirt] – has three valency members:
ACT (who is sewing)
PAT (what is being sewed)
DIR3 (directional – where the button is being sewed)

přišít² [abstractly: pin sth on sb] – *přišít mu jednu* [paste sb one] – has three other valency members:
ACT (who is pasting)
PAT (what is being pasted – slap)
ADDR (to whom it is being pasted)

In other cases the valency members for two valency frames get exactly the same functors but the semantic difference of both meanings is apparent. Also for such cases there are more valency frames in the lexicon:

(17) *dělat*: [make]
dělat¹ (be somebody/make?) – *dělat šéfa* [be a boss] – has two valency members:

ACT (who is the boss)
PAT (be whom) koho dělá?

dělat² (be concerned (with st.) – *dělat politiku* [be concerned with politics] – has two (the same) valency members:
ACT (who is concerned)
PAT (with what is concerned)

In cases where the sense distinction is not quite clear, only one valency frame has been entered in the dictionary. Thus when comparing the PDT-Vallex entries to the Wordnet (Fellbaum, 1998), which uses quite fine-grained approach to word senses, PDT-Vallex has less entries.

(18) ACT(.1) LOC() *hučet v kamíně* [whistle in the chimney]

(19) ACT(.1) LOC() *hučet v uchu* [buzzing in the ear]

(20) ACT(.1) PAT(.4) *dostat zloděje*

(21) ACT(.1) PAT(.4) *dostat vztek*

As these examples suggest, polysemy distinctions (as defined e.g. in Čermák 1995), are not completely consistent (as measured by rigorous lexicography criterions) in the PDT-Vallex; the different meanings are 'split' or kept together with a single valency frame in a rather intuitive way. Defining the type and number of meanings for a given verb is very difficult and the borderline between individual meanings is often very hard to determine, especially when they have identical information in their valency frames.

We have at least attempted to systematically distinguish among concrete (literal), abstract (figurative) and phraseological (idiomatic) semantic types even though these types are not explicitly recorded in the PDT-Vallex (for an overview and example of all types for a single verb (*nést, to carry*), see Table 1).

The three general types of meanings, considered in the PDT-Vallex, are as follows:

1. The concrete meaning of a verb: this is such a meaning that follows directly from its lexical semantics; this is the basic, non-figurative meaning.

   For example:
   ustoupit [step back] – ACT(.1) DIR1() *ustoupil od okna* [he moved away from the window]

2. The abstract (figurative) meaning: this is such a meaning that originates from using the concrete meaning in a metaphorical (figurative) sense.

   For example:
   ustoupit [step back] – ACT(.1) PAT(.3) *ustoupil výhrůžkám* [he yielded to threats]

3. The phraseological (idiomatic) meaning: this is a meaning which the verb gets when it is used as a part of a complex non-compositional lexical unit. Often, it can also be considered an abstract meaning but because of its frequent appearance, we consider it a separate type. In addition, it has two subtypes labelled CPHR and DPHR:

For example:

mít zájem [to be enterested in]
ACT(.1) CPHR({zájem,...}.4) *má zájem o koupi nového domu* [he is interested in buying a new house]

mít smrt na svědomí [lit. to have death on conscience]
ACT(.1) DPHR(na-1[svědomí.S6]) PAT(4)

One of the two subtypes of the idiomatic class concerns the semantically 'empty' meanings of verbs (sometimes called 'light verbs' in English). This meaning is used for such cases in which the meaning of a whole structure containing the verb and a noun phrase as its dependent is concentrated on the nominal part. The semantically empty meaning is indicated with the CPHR functor, assigned to the root of the nominal part. For example in the phrase *podat žalobu* [file a complaint] the main meaning is in the noun *complaint* and the verb *to file* is what we call 'semantically empty'.

Other examples:

poskytl jim pomoc.CPHR [lit. he-gave them assistance]

dostala jsem od matky pochvalu.CPHR [lit. I-have-earned from mother prise.]

The valency members of the valency frame for basic meanings are labelled with the 'common' functors (arguments as well as obligatory adjuncts). The valency members of the valency frame for abstract meanings are labelled either by common functors, or quite frequently by the CPHR or DPHR functor for multi-word complementations, since they often become idiomatic.

Nést [carry]		
Meaning	Valency Frame	Example
Basic	ACT(.1) PAT(.4) ADDR (.3)	*nést tatínkovi knihy* [carry books to father]
Abstract one-word	ACT(.1) PAT(.4)	*nést jméno* [carry name]
Abstract multi-word (idiomatic)	ACT(.1) DPHR (kůže:S4, na-1[trh:S4]	*nést kůži na trh* [lit. carry skin to market]
Semantically empty	ACT(.1) CPHR {odpovědnost,…}.4	*nést odpovědnost* [carry responsibility]

Table 1: Example: the verb *nést* [carry] with all its meanings in the PDT-Vallex:

3.3 The number of members in the valency frame

The number of members in the valency frame is fixed. Usually one valency member corresponds to one given functor. The number of functors corresponding to one valency member can be higher only in cases in which the valency frame does not consist of one given functor but of a list of alternating functors. Even in this case, the number of members in the valency frame is considered to remain fixed – which is in line with our definition of "alternating" functors, which requires that only one of them can appear in any occurrence of the frame in actual utterances (or at most one of them, if they are moreover marked as optional).

There are cases in which the valency frame does contain any members, i.e. the count is zero - no arguments, no obligatory adjuncts). This kind of valency frame is called an "empty"

valency frame and it is also labelled as EMPTY. Verbs with an empty valency frame are for example *foukat* [to blow], *hřmět* [to thunder], *rozednívat se* [dawn], *sněžit* [snow].

3.4 Labelling the members of the valency frame

The individual members of the valency frame are labelled by functors. In accordance with the principles of the FGDVT, one valency frame cannot consist of two members with the same label (i.e. the same functor).

The members can be labelled either by functors for arguments, or by functors for adjuncts (cf. Chapter 1.2; for full description, see Mikulová et al., 2005). The functor CPHR (Compound PHRaseme, or complex predicate) is assigned to such members of the valency frame which would normally be actants but because of the 'emptiness' of the verb semantics they represent the core of the meaning of the whole structure. The functor DPHR (Dependent PHRaseme) is assigned to the non-verbal part of idiomatic structures (see also Chapter 3.2).

Functor	Full term	Example
ACT	ACTor	Maminka.ACT vaří [Mother.ACT is cooking]
PAT	PATient	Jan maloval obraz.PAT [John drew a painting.PAT]
ADDR	ADDResse	Darovala Mirce.ADDR knihu [She gave a book to Mirka.ADDR]
EFF	EFFect	Přeložila knihu do angličtiny.EFF [She translated a book into English.EFF]
ORIG	ORIGin	Půjčil si peníze od kamaráda.ORIG [He borrowed money from a friend.ORIG]

Table 2: Functors assigned to arguments

CPHR	Compound PHRaseme	Pokládal otázky.CPHR [He asked questions.CPHR]
DPHR	Dependent PHRaseme	Lomil rukama.DPHR [He wrung his hands.DPHR]

Table 3: Functors assigned to valency complementations in complex predicates and idioms

In our approach we have 36 functors for verbal adjuncts in total (see below; also see Chapter 1.2 and Mikulová et al., 2005; 427).

Obligatory verbal adjuncts are always recorded in the valency frames in the PDT-Vallex. Local adjuncts (LOC, DIR3) are the most common obligatory adjuncts recorded in the valency frames.

The order of the members of a frame (based on their functors) is given only by convention (cf. Chapter 3.1, last paragraph). The canonical order is as follows: ACT, CPHR, DPHR, PAT, ADDR, ORIG, EFF, BEN, LOC, DIR1, DIR2, DIR3, TWHEN, TFRWH, TTILL, TOWH, TSIN, TFHL, MANN, MEANS, ACMP, EXT, INTT, MAT, APP, CRIT, REG.

The labelling of a member of the valency frame consists, in some cases, of several alternating functors. An example containing alternating functors can be one of the meanings of the verb *chovat se* [to behave]:

(22) chovat se [to behave]: ACT(.1) MANN(*)|CRIT(*)|ACMP(*)|BEN(*)|CPR(*)
chovat se laskavě.MANN [to behave kindly.MANN]
chovat se podle pravidel.CRIT [to behave according to the rules.CRIT]

chovat se otrocky.CPR [to behave in a slavish manner.CPR]
chovat se bezchybně.ACMP [to behave without mistakes.ACMP]
chovat se ku prospěchu věci.BEN [to behave to benefit of the issue.BEN]

Unfortunately, the situation might be more difficult to solve with this strategy if it is not clear whether the difference in two usages, each with the most obvious functor assigned, is semantically relevant or not. In such a case, two different valency frames are kept in the PDT-Vallex lexicon. This appears mostly for subtle differences in temporal or local meanings. For example, the verbs *namontovat* [to install] or *umístit* [to place] have two valency frames:

(23) ACT(.1) PAT(.4) LOC(): *namontovat zařízení v autě* [to install a gadget in the car]
(24) ACT(.1) PAT(.4) DIR3(): *namontovat zařízení do auta* [to install a gadget into the car]

(25) umístit dítě do ústavu.DIR3 [to place the child into the institution]
(26) umístit dítě v ústavu. LOC. [to place the child in the institution]

In this case, the rule "one meaning ~ (at most) one valency frame" might be compromised, but PDT-Vallex does not record the fact that these two valency frames might in fact share the same meaning.

When the real text during the PDT annotation have been confronted, it became apparent that functors not previously considered by the FGDVT might be more adequate to use in certain contexts, and that they could better distinguish semantic distinctions of the same syntactic expressions. However, given the scope of the annotation project, there was unfortunately not enough resources to explore these new functors to such an extent as to make them reliably part of both the theory and the annotation. However, some of the proposals to incorporate new functors into the FGDVT were taken at least theoretically into account (e.g. Lopatková a Panevová, 2005).

3.5 The obligatoriness of the members of the valency frame

Members of the valency frame are characterized according to their obligatoriness. (For further detailed information conf. Panevová, 1974-75). Valency frames in the PDT-Vallex contain only arguments (both obligatory and optional) and obligatory adjuncts. Optional adjuncts are not recorded in the valency frame.

3.6 Surface realization of a valency frame

It has been already shown (Hajič and Urešová, 2003) that in general the possible surface realizations of the verb, its arguments and adjuncts are dependent on the whole valency frame, and typically cannot be described purely morphematically, as it is assumed as being usual for inflective languages – complex phenomena such as agreement (which is, however, often independent of the valency frame), adjuncts, parts of some idioms and other relations sometimes complicate things in an unprecedented number of places. In fact, the aforementioned paper shows that the explicit division of the syntactic description to an analytic (surface-syntactic) one and the tectogrammatical (deeply syntactic) one is justified also by the relation between valency frames and the need to describe their surface realization explicitly. The analytic dependency trees help exactly with the difficult phenomena such as agreement, which has to be solved separately from the specifics of the particular valency frame – that is, more generally.

However, in most cases and as a basis for the more complicated ones, the relation between a particular complementation of a verb and its own surface realization (conditioned on the valency frame – i.e., roughly, the verb sense – in which it appears) can be described independently of the required verb surface form and the other complementations. In the following text, we will describe such relations, since that is also how they are used in the PDT-Vallex, ignoring the more complex relations for the moment.

The description of every complementation, by default represented syntactically and/or semantically by its unique functor within the valency frame, is extended by a fully formal description of the appropriate surface realization (surface morphosyntactic form). While the functors and their possible obligatoriness characteristics are described above, here we will concentrate on the morphosyntactic form specification.

The most important thing we should stress here that while the description of form as entered in the PDT-Vallex lexicon is relevant for active verb forms only (as it is usual in such dictionaries, even for human use), we have formal methods for transforming such a description (again fully formally and explicitly) into different morphosyntactic forms, as appropriate for various secondary diatheses in which the verb can appear (see below and in Pajas and Urešová, 2009). The morphosyntactic requirements for secondary diatheses are thus present in PDT-Vallex implicitly (see also below).

The formal means for specifying the necessary morphosyntactic form of verb complementations (or the verb itself) are however the same for any diathesis – they are in any case simply general surface-syntactic tree fragments with (only) certain fixed requirements on their shape and morphosyntactic attributes, after all.

We have defined formal means for describing both the structure of the surface-syntactic tree fragment and the requirements (also called constraints) on some of its attributes, such as lemma and certain categories in its morphosyntactic tag; rarely, but possibly also the value of surface-syntactic function (analytic function) can be formally restricted.

The structure of the tree fragment, which is either a subtree (the leaves of which can further be modified by dependent nodes, at least in general) or (rarely) a list of subtrees, is defined using a bracketing notation, where a (square) bracket denotes that the comma-separated list that follows should depend on the node immediately preceding the bracket. In other words, the notation

```
node1[node2,node3,node4[node5]]
```

says that `node1` is the direct governing node (in the surface-syntactic sense) of nodes `node2`, `node3` and `node4` (i.e., `node2`, `node3` and `node4` depend on `node1` in the surface-syntactic tree), and that `node4` is furthermore modified by `node5` (`node5` depends on `node4`). Every `node<x>` stands here for a series of morphosyntactic constraints on the node's attribute values (see below).

Each node (such as `node2`, for example) contains (typically) a short sequence of symbols that describe conditions (or in other words, restrictions) on what combination of various grammatical (morphemic and syntactic) attributes are acceptable as the surface expression for the verb complementation which it expresses. For example, if the symbols should express that the only acceptable form is a syntactically nominal expression in nominative, simply the symbol '`1`' (digit 'one') is used (morphemic cases are numbered in standard Czech grammars, from 1 to 7 – i.e., nominative to instrumental). There are no 'default' values for the relevant grammatical attributes – all restrictions must be (and in the PDT-Vallex, they are) explicitly entered in the dictionary, otherwise it would mean that there are essentially no restrictions (on top of the usual general syntactic and morphemic rules).

The description itself, in the compact form used in the PDT-Vallex, starts (from the left) with a lemma; lemmas are used mostly in the description of idiomatic expressions (the

corresponding slot is marked by the functor DPHR), light verb constructions (functor CPHR) and when requesting a particular preposition or subordinate conjunction. If the lemma is missing, then any lemma is acceptable. After the lemma, there is a separator (typically, a full stop), marking the start of the other morphosyntactic form descriptors; such a separator is necessary not only to mark the end of the lemma (unless the square bracket denoting a dependent node, or the end of the whole specification follows), but also it is indispensable to show that the lemma is missing (then it becomes the first character of the description). Please note that the lemma mentioned in this paragraph is the morphosyntactic lemma, and (exceptions notwithstanding) it should also contain the morphological lemma identification number in case it has such an identifier assigned in the morphological dictionary (e.g. '`stát-3`' for '*to_stand*', '`stát-1`' for '*a_state*'). Some examples:

(27) `proti-1` the lemma for preposition `against'

(28) `proud.` the noun 'stream', followed by the lemma/morphosyntax separator

The morphosyntactic part starts with an optional, single-character symbol most closely described as a part-of-speech descriptor. This descriptor, however, might in fact mean more than a mere restriction on a part of speech of the surface expression – generally, it corresponds to a number of restrictions on attributes of the surface-syntactic tree node. Whereas the symbols '`a`' (adjective), '`n`' (noun), '`d`' (adverb), '`i`' (particle), '`u`' (possessive pronoun), '`v`' (verb) and '`j`' (subordinate conjunction) translate essentially to a simple restriction on the morphosyntactic part of speech, some others should be viewed as a shorthand for one or more additional restrictions. For example, '`f`' means verb which is additionally in infinitive, '`s`' means a node which is the root of a subtree marked as direct speech, '`c`' means verb as a root of a subtree containing a surface dependency representation of a relative clause introduced by a relative pronoun or adverb. When missing, any part of speech is allowed unless restricted by other morphosyntactic descriptors that follow or the preceding lemma. Again, some examples:

(29) `že[.v]` a verb dependent (in the surface representation) on the conjunction '`že`' ('*that*') Note: it is unnecessary to put '`že.j`' since the lemma 'že' is only a '`.j`'

(30) `jazyk[.u]` the lemma '`jazyk`' ('tongue') modified by a dependent which must be a possessive pronoun (as in '*rozvázat něčí jazyk*', lit. '*untie sb.'s tongue*')

The (compact) part-of-speech designation is immediately followed by gender descriptor (if any), which resembles the true morphosyntactic gender tag symbol but in fact is a bit more general and corresponds directly to the traditional notion of gender in Czech. There are four such genders being distinguished in Czech: masculine animate (symbol used: '`M`'), masculine inanimate ('`I`'), feminine ('`F`') and neuter ('`N`'). These symbols then 'expand' to match other symbols describing some gender alternatives (sets), such as '`H`' (for neuter or feminine), '`Y`' (masculine, regardless of animateness), '`Z`' (anything but feminine), etc. While the gender descriptor is almost unused (since gender is driven almost exclusively by surface-syntactic agreement rules, and never conditioned on the functor), the next one, the grammatical number, is not infrequent, especially when the surface expression is restricted to certain lemmas in the given grammatical number only. As expected, the symbols for number are '`S`' for singular and '`P`' for plural and dual, leaving the true 'dual' ('`D`') for grammatical agreement rules to sort out, if ever needed. Before giving combined examples, the most frequent descriptor should be introduced, namely, the morphemic case, which we have briefly mentioned above.

Morphemic case is, as opposed to gender and number, very often determined and restricted very narrowly (typically, to a single allowed morphological case) by the verb and the particular verb complementation slot name. In Czech grammars, cases are numbered: '`1`' for

nominative, '2' for genitive, '3' for dative, '4' for accusative, '5' for vocative, '6' for locative and '7' for instrumental. It is exactly these symbols which are used for specifying this important restriction. However, these symbols also allow for numerical expressions actually in the form of a surface-syntactic tree for an expression that contains some numbers related to the nominal phrase itself; for example, in 'viděl méně než pět lidí' (lit. '*he-saw less than five people*') the phrase 'méně než pět lidí' also corresponds to '4', even though it is in fact a subtree rooted in 'méně' (which is a case less adverb) with the numeral 'pět' (in the requested accusative case) hidden somewhere down the tree.

Some examples for number and/or case descriptors:

(31) `4` accusative case, such as in 'vidět ho' ('*see him*')

(32) `2` genitive case ('týká se ho to', lit. '*concerns itself him it*')

(33) `o-1[.6]` preposition '`o`' ('*about*') and (as a dependent in the surface-syntactic dependency tree) a locative nominal expression

(34) `za-1[ucho.P4]` in 'tahat za uši' (lit. '*pull for ears*', '*musically unpleasant*')

So far, the symbols for part of speech, gender, number and case are mutually disjoint and therefore no separators are needed in the description. For degrees of comparison (adjectives and adverbs only), the symbol '`@`' precedes the descriptor of degree restriction (expressed again as a number, '`1`', '`2`' or '`3`' for positive, comparative and superlative, respectively). The last special symbol in the string of descriptors is (again optional, as all the attribute-restricting symbols described so far are) '`#`', a request for the usual gender, number and case agreement with the immediately preceding governing node, which normally would not be detected by the general grammatical rules. Examples:

(35) `na-1[bedra.6[.u#]]` in 'leží to na jeho bedrech' (lit. '*lies it on his shoulders*')

'bedra' must be in locative, and modified by a possessive pronoun agreeing with 'bedra' in number (plural), gender (feminine) and case (locative).

If the above descriptors are insufficient to accurately describe all the necessary restrictions on the surface form of the node, it is also possible to encode them by using direct reference to the individual positions of the morphosyntactic tag as used in the morphological analyze used in the Prague Dependency Treebank (Hajič, 1994). The position of the symbol (i.e., the category number) is introduced by the symbol '`$`', and this "index" is then followed by a single-character symbol enclosed in '`<`' and '`>`' (e.g., `$14<6>` requests the form be colloquial in its ending). All the symbols within the angle brackets must be escaped by a backlash unless alphanumeric. Finally, negation of the form can be requested by a tilde ('`~`').

It is not uncommon that a given verb complementation can in fact be expressed in several different forms. The most typical alternative is a simple case (direct object, for example) and one or more prepositional case(s) (indirect object(s)). Each alternative is then composed as described above; they are separated by a semicolon ('`;`').

For (obligatory) adjuncts, the complete form designation can be left out if it corresponds to the typical set of allowed forms for the given functor. For example, LOC is expressed by a wide range of prepositional constructions, which typically are all valid regardless of the verb and its sense (as opposed to arguments, or adjuncts in our terminology, where the form must always be entered, even if it is the "usual" one).

This chapter will conclude with a set of examples, from simple to more complex ones, to illustrate the possibilities of describing restrictions of form. Please note that the link between

the verb argument slot (which is itself described by a functor) is established by simply enclosing the form description by parentheses immediately after the functor name, e.g., the specification PAT(`.4`) associates the argument named PAT (of the given PDT-Vallex entry) with the request that it be in accusative case (with no further restrictions).

Form of surface expression examples:

(36) Restriction on morphological case only: `.4`

(37) Preposition and case: `s-1[.7]`

(38) Preposition and a case, or simply a case (both forms allowed): `pro[.4];.3`

(39) Dependent clause (root is a verb) introduced by the subordinate conjunction 'že' ('*that*') or 'aby' ('*to*'): `že[.v];aby[.v]`

Full valency frame description examples:

(40) Standard transitive verb frame: ACT(`.1`) PAT(`.4`)

(41) Infinitival expression for deep object: ACT(`.1`) PAT(`.f`)

(42) Phraseme ('běhat', lit. '*run*'): ACT(`.3`) DPHR(`mráz.S1,po-1[záda:P6]`)

(43) Optional arguments: ACT(`.1`) PAT(`.4`) ?ORIG(`z-1[.2]`) ?EFF(`na-1[.4]`)

4 The relation between PDT-Vallex and the Prague Dependency Treebank

The PDT-Vallex dictionary has been conceived to be an integral part of the Prague Dependency Treebank (PDT). More precisely, it has been designed to be interconnected with the tectogrammatical annotation of the texts contained in the PDT. This "interconnection" is implemented as links that point from every occurrence of a verb (and valency-bearing nouns) to the appropriate entry in the PDT-Vallex (see Chapter 3).

While there are other corpora which do include sense distinction and even verb or noun valency (albeit under a different name, most notably the Penn Treebank and PropBank/NomBank, see Palmer and Kingsbury, 2005, Meyers, 2004a, 2004b, Marcus et al., 1993), one feature which is unique for the Prague Dependency Treebank is that the surface form specification, as entered in the PDT-Vallex and described in the previous chapter, has actually been fully formally and automatically cross-checked against the treebank annotation.

Obviously, this has been an iterative process, when all mismatches found during this automatic check have been subsequently manually checked in both places – in the PDT-Vallex and in the PDT data, and corrected appropriately either by correcting the PDT-Vallex entry, adding a new one, etc., or by correcting the data annotation in the PDT.

Since the PDT-Vallex contains only the surface form specification for active voice (primary diathesis), it is apparent that there would be a mismatch for every secondary diathesis use in the corpus. Therefore, a set of automatic transformation rules has been created and implemented (Pajas and Urešová, 2009). These rules take the surface form specification as found in the PDT-Vallex for the primary diathesis, convert them to a form corresponding to the secondary diathesis used in the data (where the particular verb sense occurs) and check if that "transformed" form is present in the data, informing of any remaining discrepancies. The secondary diatheses found in the PDT data and handled by the transformation rules cover reflexive passivization, periphrastic passivization, resultative, disposition modality and reciprocity.

Thus in the cross-checking iterations, these transformation rules for diatheses have also been gradually built and improved as the PDT data annotation and PDT-Vallex entries were getting closer and closer to a mutually consistent state.

5 Conclusions

Primarily, the PDT-Vallex served for keeping inter-annotator consistency high during the process of manual corpus annotation, most importantly for functor assignment to verbal complementations. After the tectogrammatical annotation process has ended, the lexicon served also for rigorous, automatic cross-checking of the annotated PDT data against this newly built lexicon. The PDT-Vallex is publicly available as an integral part of the PDT version 2 published by the Linguistic Data Consortium.[5]

Notes

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

2 For detailed information see our web pages: http://ufal.mff.cuni.cz/pdt2.0/doc

3 The „standard" FGDVT was later enriched, i.a., through so-called quasi-valency and typical valency modifications. See (Lopatková and Panevová, 2005).

4 By 'diatheses' we mean the relation between semantics of the verb and its surface realization.

5 http://www.ldc.upenn.edu, Catalogue number LDC2006T01

Acknowledgements

The research reported in this paper was supported by the Grant Agency of the Charles University in Prague No. GAUK 52408/2008, and Czech Republic Ministry of Education Projects Nos ME09008 and MSM0021620838.

References

Čermák, F. (1995). *Manuál lexikografie*. H+H. Praha.

Daneš, F. (1985). *Věta a text. Studie ze syntaxe spisovné češtiny*. Academia, Praha.

Fellbaum, Ch. (1998). *WordNet: An Electronic Lexical Database*. 445p. Cambridge, MA and London. MIT Press.

Hajič, J. (1994). *Unification Morphology Grammar.* PhD Dissertation. MFF UK, Charles University, Prague.

Hajič, J. and Z. Urešová (2003). Linguistic Annotation: from Links to Cross-Layer Lexicons**.** In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, 69-80. Vaxjo University Press.

Hajič, J., J. Panevová, E. Buráňová, Z. Urešová, A. Bémová, J. Kárník, J. Štěpánek, P. Pajas (1999), *Annotation at Analystical Level.* Annotation Guidelines. Available online at http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/html/index.html. Accessed Sep. 26, 2009. Additional documentation online at http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/en/html/ch05.html. Accessed Sep. 26, 2009.

Hajič, J., J. Panevová, Z. Urešová, A. Bémová and V. Kolářová. (2003). PDT-Vallex: Creating a Large-coverage Valency Lexicon for Treebank Annotation**.** In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, 57-68. Vaxjo University Press.

Lopatková,M. and J. Panevová (2005). Recent developments of the theory of valency in the light of the Prague Dependency Treebank. In: *Insight into Slovak and Czech Corpus Linguistics*. Mária Šimková, ed. 83-92. Veda Bratislava, Slovakia.

Meyers, A., R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young and R. Grishman. (2004). The NomBank Project: An Interim Report. *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, Boston, MA, 24-31.

Meyers, A., R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young and R. Grishman. (2004). Annotating Noun Argument Structure for NomBank. Proc. *Fourth Int'l Conf. on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

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

Mikulová, M., A. Bémová, J. Hajič, E. Hajičová, J. Havelka, V. Kolářová, M. Lopatková, P. Pajas, J. Panevová, M. Razímová, P. Sgall, J. Štěpánek, Z. Urešová, K.Veselá, Z. Žabokrtský and L. Kučová. (2005). *Anotace na tektogramatické rovině Pražského závislostního korpusu*. Anotátorská příručka. TR-2005-28. Prague. ÚFAL MFF UK, Prague. Available online also in English at http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/index.html, and as a downloadable document at http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/pdf/t-man-en.pdf. Short reference is available in Czech and English at http://ufal.mff.cuni.cz/pdt2.0update. Accessed Sep. 26, 2009.

Pajas, P. and Z. Urešová.(2009). Diatheses in the Czech Valency lexicon PDT-Vallex, In Proceedings of *Slovko*, Nov. 25-27, Smolenice, Slovakia. in print.

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

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

Panevová, J. (1998). Ještě k teorii valence. *Slovo a slovesnost* 59, č. 1.

Panevová, J. (1999). Valence a její univerzální a specifické projevy. In Z. Hladká, P. Karlík(ed.) Čeština - univerzália a specifika. Sborník konference ve Šlapanicích u Brna.

Panevová, J. (2002). K valenci substantiv (s ohledem na jejich derivaci). In *Zbornik matice srpske za slavistiku*. 61. 29-36. Novi Sad.

Pauliny, E. (1943). *Štruktúra slovenského slovesa*, SAVU, Bratislava

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

Urešová, Z. (2005) Verbal Valency in the Prague Dependency Treebank from the Annotator's Point of View. In M.Šimková (ed.) *Insight into Slovak and Czech Corpus Linguistics*. 93-112. Veda Bratislava, Slovakia.