Diatheses in the Czech Valency Lexicon PDT-Vallex

Zdeňka Urešová and Petr Pajas

Charles University in Prague, UFAL MFF UK

Abstract. An important design element in all lexicons, whether human-oriented or designed for computer processing, is the variability of forms in which lexical units described in the lexicon entries can occur in natural language utterances. If all such forms and variations were to be listed independently in the lexicon, its size would be enormous and it would be hard to maintain (every change would have to be copied to many entries). These problems can even multiply in the case of lexicons for computerized natural language applications, where entries must be explicitly and formally described in full detail.

As an inherent part of the Prague Dependency Treebank project ([9]; for its theoretical background, see the work of Sgall et al. [33]) a valency lexicon called PDT-Vallex ([10], [39], [40]) has been created and is publicly available, with over 8800 verb senses and their corresponding valency frames, linked fully to the treebank.

When a particular verb sense is used in a diathetic expression (passive construction, reciprocity, resultative or dispositional modality etc.), the surface expression of verb complements also changes ([40]). While the basic form "transformations" are well known, it is less obvious how to describe them for all the modalities, especially for the purposes of computer processing, where everything must be explicitly stated. We have found that these transformations can be described by a set of rules, which then allow to keep only a canonical (i.e., the active-voice) valency frame in the lexicon entry, and use these rules to obtain surface expression constraints for all the diatheses covered. This formalization have been used in the formal checking of the Prague Dependency Treebank project and it is used in other current projects as well.

1 Valency

Before we will concentrate on diatheses in the PDT-Vallex, let us make a little digression into the very notion of valency and diathesis.

1.1 Valency in general

This introductory section reviews some very basic facts about valency. Most writers on the subject cite Tesnière [38] as the one responsible for introducing the term of valency into modern linguistics. Tesnière uses the term valency for syntactic analysis of a sentence, so it was linked also to dependency. Active valency and passive valency are occasionally distinguished in literature ([22]).[1]

[1] In this article, whenever we simply refer to valency, we mean "active" valency.

After the first presentation of valency by Tesnière, the study of valency was taken up by many scholars, with a wealth of material now available. Since individual authors see valency from different perspectives, so far no generally accepted definition of valency exists (Storrer [36]). Generally, valency is understood as a specific ability of certain lexical units - primarily of verbs - to open free slots for filling in by other lexical units. By filling these slots a sentence structure is being built. Valency is seen as both syntactic, semantic, or some combination of them.

The valency terminology is also inconsistent; terms like valence, subcategorization, intention (in [30]), government, government pattern ([20]), complex sentence pattern ([4]), argument structure ([26]), stereotypical syntagmatic patterns ([32]) etc. emerge. Naturally, these terms not always denote exactly the same linguistic phenomenon. For a detailed survey see [40].

1.2 Valency in the Functional Generative Description

Among theories combining the syntactic and semantic approach, the valency theory developed within the framework of the Functional Generative Description (FGD) is found. (see e.g. [28], [29] and [16]). It uses syntactic as well as semantic criteria to identify verbal complementations. In this theory, it is assumed that potentially every (semantic) verb, noun, adjective and adverb (i.e. every complex node) has subcategorization requirements, expressed by its valency frame. Valency modifications include all kinds of elements (dependency relations) that can modify a particular lexical unit.

For example, in the sentence *Jitka mu daruje knihu* (lit. *Jitka gives him a book*) the verb *darovat* (lit. *to give*) opens a slot for a subject in nominative, i.e. for an agent (*Jitka*), then a slot for a dative object, i.e. for an addressee of giving (*mu*, lit. *him*), and lastly a slot for an accusative object, i. e. for an object which is being given (*knihu*, lit. *a book*).

Since the FGD does not work with the notion of "semantic roles" as known from some of the literature (such as a "runner", "giver", "object-given", see e.g. [15]), the appropriate lexical unit (here the verb) therefore determines—besides the morphological requirements on arguments—also their semantic properties.

In FGD, we work with TECTOGRAMMATICAL REPRESENTATION of sentences, which reflect their underlying syntax and certain types of semantic attributes. In this formalism, the central position in a sentence (or clause) is occupied by a (typically) finite verb.[2].

In order to sort out the behavior of all word modifications and in order to describe their character we define the following main basic principles in our valency concept:

- a valency frame is assigned to each verb sense separately,[3]

[2] Also some nouns, adjective and adverbs valency frames are recorded in the PDT-Vallex, but we don't discuss them in this contribution.

[3] Verb senses are defined rather coarsely, as opposed to some other approaches, such as the famous WordNet resource. However, it is not excluded that two clearly distinct senses carry identical valency frames. In other words, senses are not forced to be merged just because their valency frames are the same.

- criterion for distinguishing inner participants (arguments) and free modifications (adjuncts),
- criterion for distinguishing obligatory and optional modifications, and
- the concept of "argument shifting".

According to the type of dependency, any modification can be classified as either an INNER PARTICIPANT (that is, an argument) or as a FREE MODIFICATION (which is close to what is known as an adjunct). A given modification of a particular lexical unit may be—with respect to its particular governing word—either OBLIGATORY (that is, obligatorily present in the deep, tectogrammatical structure) or OPTIONAL (that is, not necessarily present). For the obligatory vs. optional distinction, the DIALOG TEST, described later, is used.

The distinction between inner participants and free modifications is *not* verb-specific: if a dependency type is an inner participant, then it is considered an inner participant for all verbs which it possibly modifies. We have determined that there are five such types of arguments: actor (ACT), patient(PAT), addressee (ADDR), effect (EFF), and origin (ORIG). These arguments have also the additional property that they can appear only once in a given clause headed by the verb (in the particular verb sense) to which they belong.

Among the 70 complementation types used in the Prague Dependency Treebank, we identify 36 verb free modification types (adjuncts): adjuncts expressing semantic time relations: TFHL, THL, THO , TFRWH, TOWH, TPAR, TSIN, TTILL a TWHEN; adjuncts for local semantic relations: DIR1, DIR2, DIR3 a LOC; adjuncts for causative relations: ACMP AIM, CAUS, CNCS, COND a INTT; adjuncts for means relations: CPR, CRIT, DIFF, EXT, MANN, MEANS, REG, RESL a RESTR; modal adjuncts: ATT, INTF a MOD; semantically different adjuncts: BEN, CONTRD, HER a SUBS, and finally, adjuncts with double semantic dependency (verb and another verb agrument): COMPL.

While arguments can modify just a relatively closed class of verbs, every adjunct can modify (in principle) any verb. That is also where their name (free modification) comes from; moreover, they can be repeated within a given clause several times.

For distinguishing among the five inner participants we use syntactic as well as semantic criteria. Actor (ACT) is always the first inner participant (something like Arg0 in the PropBank) and Patient (PAT) is always the second inner participant (usually like Arg1 in the PropBank, see [26]). These two arguments are thus determined more or less syntactically. Only when a verb has more than two arguments, semantic criteria come into play. Semantic origin (for example, *to make of wood*) gets the label ORIG, semantic addressee (*talk to somebody*) gets ADDR and semantic result (effect) gets the label EFF (*to split into pieces*).

To stress the distinction between the typology of the first two arguments of the verb and the rest (if any), FGD has adopted the concept of shifting of arguments. According to this special rule, semantic Effect, semantic Addressee and semantic Origin (which would normally be labeled by EFF, ADDR and ORIG, respectively) 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 labeled PAT (as it is in all sentences headed by the same sense of the verb *to dig*); similarly, in the sentence *The* *teacher asked the pupil*, the semantic Addressee is shifted to the Patient position. This rule simply helps to keep consistency at the expense of lower semantic "precision".

Both arguments and adjuncts can be in their relation to a particular word either obligatory (i.e., obligatorily present at the tectogrammatical level of sentence representation) or optional (i.e. not necessarily present in each sentence where the verb is used). It should be stressed that this does not concern the surface appearance of such modifications, because they can be elided virtually anywhere; the notion of obligatoriness is used in the semantic sense. A natural question arises how the obligatoriness can then be determined, given that surface appearance cannot be used as a criterion: we rely on a DIALOG TEST [27]. The dialog test 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 the speaker has used. If the speaker can sensibly answer a hearer's follow-up question about a semantically obligatory modification "I don't know", then it means that the given modification is optional. On the contrary, if the answer "I don't know" is not possible in the particular point of this dialog-to-be, then the given modification is considered obligatory. For example, if the verb *to leave* (in the sense of "departing") is used in a sentence *John left*, the speaker must know from where John left (otherwise, he or she would—even should—have used another verb). Consequently, "from where" (DIR1) is an obligatory modification. Conversely, the speaker does not need to know to where John left—thus, if present, the "to where" (DIR3) modification will always be optional.

1.3 Valency and the Prague Dependency Treebank

The concept of the valency frames in the Prague Dependency Treebank (PDT) annotation ([21]) corresponds to the valency theory built in the FGD framework described above.

The work on the valency lexicon enabled the confrontation of the valency theory and real usage of language. Thus, we can say that PDT-Vallex has been created "bottom-up"; it was not necessary to make up valency complementation examples for the theoretically given schemes of valency frames because the lexicon draw upon the real texts from a real corpus.

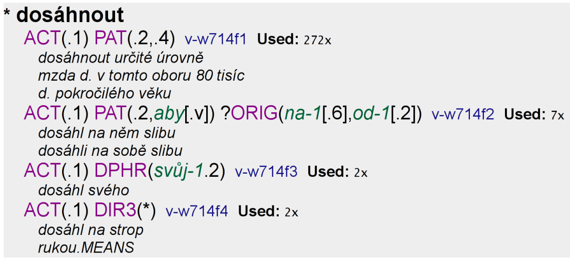
Primarily, the PDT-Vallex served for keeping inter-annotator consistency high during the process of manual corpus annotation, most importantly for functor assignment to verbal complementations. After the tectogrammatical annotation process has ended, the lexicon served also for rigorous, automatic cross-checking of the annotated PDT data against this newly built lexicon.

The PDT-Vallex contains only those words (verbs, nouns, adjectives and adverbs) and their senses which occurred in the annotated data. The lexicon contains 10039 different words: 5510 verbs, 3727 nouns, and a small number of adjectives and adverbs. The total number of valency frames is 14979, out of which there are 8810 valency frames for verbs.

The valency modifications are described in the valency frame of the particular verb. Arguments (inner participants) are always recorded, be they obligatory or optional; adjuncts (free modifications) are recorded only if determined obligatory.

Apart from the obligatoriness indication we also record the dependency relations (the FUNCTOR) and the morphemic surface form. Every verb has at least one valency frame; each frame corresponds to one sense (meaning) of the verb.[4]

For example, the (English) verb *to leave*, which has two clearly distinguishable senses, would have two valency frames in our valency lexicon. The first one would be used for the sense *somebody left something* (with an Actor and a Patient as the two obligatory arguments) and the second one for the sense *somebody left from somewhere* (with and Actor and Direction-from as the obligatory arguments for this sense).



```
* dosáhnout
  ACT(.1) PAT(.2,.4)  v-w714f1  Used: 272x
    dosáhnout určité úrovně
    mzda d. v tomto oboru 80 tisíc
    d. pokročilého věku
  ACT(.1) PAT(.2,aby[.v]) ?ORIG(na-1[.6],od-1[.2])  v-w714f2  Used: 7x
    dosáhl na něm slibu
    dosáhli na sobě slibu
  ACT(.1) DPHR(svůj-1.2)  v-w714f3  Used: 2x
    dosáhl svého
  ACT(.1) DIR3(*)  v-w714f4  Used: 2x
    dosáhl na strop
    rukou.MEANS
```

Fig. 1. The PDT-Vallex entry for *dosáhnout* (*to reach*)

A real PDT-Vallex entry for the verb *dosáhnout* (*to reach*) can be seen, formatted for better readability, in Figure 1. This verb has four different senses in our dictionary. The first sense *to reach something* corresponds to the first frame, which contains ACT in nominative and a PAT in genitive or in the accusative morphemic case. This frame has been used 272 times in the data. Below the formal description of the valency frame, three usage examples can be found. Similarly, the other three senses are described using the same structure.

An example sentence with the verb *dosáhnout* used as the main verb is in Figure 2. Obviously, this is an example of the usage of the most frequent sense (first entry as seen in Figure 1). The Actor (ACT) is the word *kurs* (*price*), and the other obligatory argument is the Patient (PAT)—the word *hodnota* (*value*), further modified by the actual price tag and the currency designation.

In Figure 3, we have schematically depicted how the corpus is linked to the PDT-Vallex lexicon. Let's say that in the corpus, we have 3 occurrences of the verb *uzavřít* (lit. *to close*) in 3 sentences. There are two PDT-Vallex entries (valency frames) for *uzavřít*. 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. The third occurrence of *close* is linked with the first valency frame, which represents the light verb meaning, denoted here with the CPHR functor in its frame.

[4] In rare cases, the description of which is beyond the scope of this article, two or more valency frames may still be used with the same sense of the verb. This is however not reflected in the current format of PDT-Vallex. For a suggestion of possible restructuring to allow for (i.a.) such grouping, see [42].

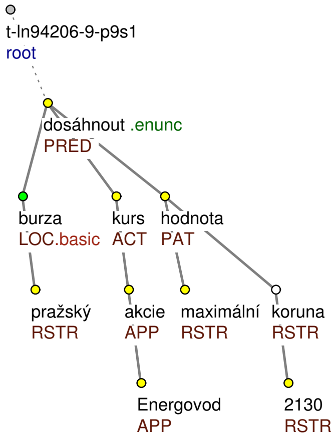


Fig. 2. The (simplified tectogrammatical representation of the) sentence *Na pražské burze dosáhl kurs akcií Energovodu maximální hodnoty 2130 korun.* (Lit. *On the Prague Stock Exchange reached the price of shares of Energovod the maximum value of 2130 Czech crowns*)

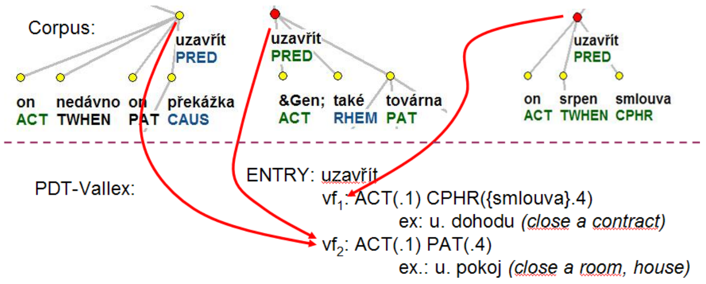


Fig. 3. Links between the corpus and the PDT-Vallex entries

The arguments are linked implicitly, and their correctness in form can be determined using the valency entries and the diathesis transformation rules described in this article.

A detailed description of the surface form of a valency modification as captured in the valency frames in the PDT-Vallex can be found in Sect. 3.1 of this article, because this information is relevant for the description of diatheses and their surface transformation rules; for the most detailed information, see the annotation manual ([21]).

2 Diathesis

2.1 Diathesis in general

In contemporary linguistics, the very term DIATHESIS is closely related to other terms, e.g. alternation ([15]), conversion ([1]), hierarchization ([5]) or genus verbi ([12]). The phenomenon of diathesis was elaborated in detail in the aforementioned work of F. Daneš and in other books and articles ([4], [35], [6], among others). Lately, this question is researched in relation to the valency lexicon [14]. Worldwide, the diathesis issue is elaborated e.g. by Mel'čuk [19], Chrakovskij [3], Padučeva [24] and [23], Uspenskij [41] or Babby [2].

The DIATHESIS is understood as a syntactic grammatical category related to verb voice. It is defined, e.g., as the "relation among the elements of the semantic structure of the sentence and their corresponding syntactic positions" ([13], p. 522).

A given proposition can appear in the syntactic structure of the utterance either in its primary (basic, or canonical) diathesis and in a number of secondary (derived) diatheses. The primary diathesis is defined as the use of the verb in active voice (and in finite form). More general definition (12) says that the primary diathesis is the one which co-occurs with the highest number of complements on the surface, with the subject to the left of the verb.[5] Other diatheses, in which the semantic (deep) subject (or ACT in our case) is *not* in the subject position in the surface structure, are considered secondary. In Czech, the secondary diatheses are signaled by specific verb forms (such as the passive voice) and syntactic structures which force the semantic subject to move out of the surface subject position.

Specifically, periphrastic or reflexive passive constructions, constructions with the verb "to have" with verbal passive participle (resultative, similar to perfect tense in English), dispositional modality constructions, or the construction "to get" with a passive participle (causative) are all examples of secondary diatheses in Czech.

Sometimes, alternations are also classified as diatheses, such as in Kettnerová and Lopatková [14], where they distinguish grammatical diatheses (roughly, the ones mentioned in the previous paragraph) and semantic diatheses (alternations as the term is defined elsewhere).

The Czech valency dictionaries, both printed ([37], [17]) and electronic ([18], [25], [11]), contain (if ever) just canonical forms of the verb complementations used in the primary diathesis. The only exception is the dictionary of Skoumalová [34], who captures also the explicit complementation forms used in the secondary diathesis. However, for the natural language processing (as well as for verifying the theoretical language description) a valency lexicon with a systematical description of all syntactic and morphosyntactic forms is needed; the analysis and synthesis of Czech language, i.e. information about the morphematic realization of particular verb complementations, would be helpless without this piece of information.

3 Transformation Rules for Diathesis in PDT-Vallex

In this section, we first describe the means for formalizing form of expression of the verb and especially its arguments in PDT-Vallex. Then, after describing the general ideas behind formalizing form transformation occurring in diatheses, we describe the types of diatheses that we have dealt with in PDT-Vallex in more detail and give some examples.

[5] In Czech, this implies the standard word order, i.e. with the subject not being the focus of the sentence.

3.1 Explicit description of surface form in PDT-Vallex

As it has been said already, the PDT-Vallex entries describe, in a fully formalized way, the canonical (primary diathesis) expression of the verb and its arguments. However, with only a few extensions, the same formalism can be used for the explicit description of the necessary form of expression of secondary diatheses (i.e., the result of a transformation of the canonical surface form to the secondary diathesis one).

In all cases, the description of form should in general describe the predicate and its arguments as a whole, due to various possible interdependencies among the expression of form for the verb and its individual arguments ([8])) . However, with only a few exceptions, the form is independent among the arguments of the verb. Therefore we decided that the description of form will be associated with the individual arguments, independently of each other. This is true for both the canonical form (as present in PDT-Vallex) as well as for the resulting transformed form descriptions for the diatheses, even though the transformation rules themselves have to consider several or all the arguments at once.

The part-of-speech and morphosyntactic requirements which characterize the surface form corresponding to the valency slot (argument) in question are denoted by a short, formally defined string of symbols, separated from the m-lemma (if any) by a separator (mostly a period, (.)). They appear in the following order: part of speech requirement, then the morphosyntactic requirements (values) of gender, number, case, degree of comparison, agreement and negation. If any of these designators is missing, any value of the given category is allowed (in most cases, that means it is not really relevant for the relation between the verb and the argument in the corresponding slot).

The first designator (for part of speech) sometimes carries some additional requirement, such (for a verb) to be an infinitive. At the part-of-speech position, clausal restrictions (if the complement is realized as a clause, and not as a noun or other simple phrase). Lowercase letters are used at this position:

a	adjective
d	adverb
n	noun
i	interjection
v	verb
f	verb in infinitive
u	possessive pronoun or adjective
j	subordinate conjunction (with a clause it governs)
s	direct speech (root of subtree)
c	relative clause (root of subtree)

Gender is denoted by the following four capital letters: F for feminine, M for masculine animate, I for masculine inanimate and N for neuter.

Number is denoted by uppercase S and P with the obvious meanings.

For (morphosyntactic) case, digits are traditionally used for the seven cases in Czech: 1 for nominative, 2 for genitive, 3 for dative, 4 for accusative, 5 for vocative, 6 for locative and 7 for instrumental. The degree of comparison also uses the digits 1 to 3 for positive, comparative and superlative, respectively, preceded by the symbol @ to distinguish them from the case markers.

Agreement in gender, number and case with the governing node at the surface layer is denoted by #.

For negation, we use the tilde character (~).

In addition, any combination of the morphological attribute values that are used at the morphological and analytical layers of the PDT can be included as a requirement. Special marking separates them from the above shorthands: a $ symbol and a tag index in < and > must precede the concrete value, which then should match directly the tag position at the given index.

The surface form designation might have alternatives (separated by a semicolon), or even be empty. Empty form designation is allowed only for free modifications, and it means, that any form that is associated with the particular functor can be used.

Examples of the designation of requirements on the surface realization, roughly sorted by frequency of appearance in the PDT Vallex dictionary:

1. Case-only requirement: .4
2. Preposition and a particular case: s[.7]
3. Alternative surface expression: preposition (with a particular case), or a case-only designation: pro[.4];.3
4. A particular subordinate conjunction (alternative of two) governing a verbal clause on the surface: že[.v];aby[.v]
5. Dependent clause, no conjunction: .v
6. Multiword preposition with genitive: od-1[na-1,rozdíl,.2]
7. Phraseme *balit fidlátka* (lit. *pack one's belongings*, i.e. *to leave*): fidlátko.P4

It should be also noted that in reality, the form designators described above are rather short abbreviations for sometimes much more complicated logical expressions; for example, .1 and .4 are matched also by various prepositional or nominal forms of numerical expressions not necessarily in nominative or accusative.

In addition, a preposition requiring a single case can be abbreaviated as <preposition>+<case> (e.g., as in od+2, corresponding to the less readable format od[.2]). For the purpose of brevity, we similarly introduce (in this paper only) a single number (for example, 4) to mean a non-prepostitional, direct object in the given case (normally written as .4).[6]

3.2 Diatheses and form transformation

Again, the PDT-Vallex dictionary [39], contains only the canonical surface form designation, i.e. the one which describes the form of the verb complements in the primary diathesis appearance (active voice, finite form). This seemingly causes inconsistency with the corpus annotation, since the form as required by the form designator in the valency frame pointed to by the occurrence of the verb sense in the corpus does not match the actual form of the complements in the corpus if any of the secondary diatheses is used. However, since the change of the form when the verb appears in a secondary

[6] This latter abbreviation is not used in the real data, however.

diathesis does not depend on the particular verb or verb sense, we can use quite general "transformation rules", which can convert the form designation present at the canonical valency to its "secondary" designation which then should display a perfect match with the annotated occurrence in the corpus.

This has allowed for a complete verb sense and valency annotation of every occurrence of every verb in the corpus, while still being able to check for the correct relation between the verbal frame in the PDT-Vallex dictionary and its surface realization in the corpus even if the verb is not used in the primary diathesis form.

The transformation rules were prepared with annotation consistency checking in mind. Therefore, they aim at such transformation of the valency frame (using also the annotation information from the corpus) to arrive at a simple set of checks to either confirm or reject whether the annotation of the verb and its dependents and its context is consistent with the requirements found in the valency frame.

Every transformation rule has two parts:

1. condition to be fulfilled at the node being checked, and
2. a set of rewriting rules.

Every rewriting rule has three parts:

1. the type of the rule (replacement, alternative)
2. assertions about the verb frame
3. specification of the necessary changes in the valency frame

While we will not fully dissect all the rules in the following sections describing the individual diatheses according to the above structure, we will aim to describe all the main aspects of the transformation. Full details can be found in [40].

3.3 Transformation rules for periphrastic passivization

Only transitive verbs (i.e., verbs with a slot marked PAT in the PDT-Vallex) can appear in the periphrastic passivization type of diathesis.

The verb itself must be in the form of passive participle, and the actor (ACT) is moved from the subject position to either an instrumental-case object position (corresponding roughly to the English prepositional phrase with the preposition *by*) or it is realized as a prepositional phrase with genitive preposition *od* (lit. *from*). Sometimes, both forms are allowed ((7) as well as (od+2)). This transformation of form always takes place, regardless of what other complement gets to the subject position.

The (surface) subject can either be missing (zero pronoun form), or in fact any of the arguments can get to that position:

PAT *The painter painted a picture*.PAT
→ *A picture was painted by a painter.*
ADDR *The injury slowed down the athlete*.ADDR
→ *The athlete was slowed down by an injury.*
EFF *The teacher has read a resume*.EFF *about him*
→ *A resume has been read about him by the teacher.*

Periphrastic diathesis, if used with a perfective verb, can be also considered a resultative diathesis, if it describes a completed event. However, the aim of the use has no reflection on the form changes that the arguments undergo in this diathesis, therefore we are not making this distinction here.

In periphrastic passivization, which is by far the most frequent of all diatheses, the following cases are covered by our rules:

1. PAT is moved to the subject position; in this case, we have to further look at the form of the PAT actually found in the data. The individual forms will be transformed as follows:
   - (4) → (1): nominative case for the "moved" subject
   - (f) → (f): infinitive stays as such; verb agreement: 3rd. pers. Sg. Neuter
   - (c) → (c): relative clauses do not change; 3rd. pers. Sg. Neuter
   - other form of PAT → DELETE (i.e., does not appear on the surface)
   - other arguments: forms kept as they are recorded in the canonical frame, except if EFF(jako+4) (lit. *as* with accusative) is present in the valency frame together with PAT(4) → PAT(1), then it is changed to EFF(jako+1).
2. ADDR is moved to the subject position
   - (4) → (1): nominative case for the "moved" subject; for other forms of expression in the canonical form, no change.
   - other arguments: forms kept as they are recorded in the canonical frame
3. EFF is moved to the subject position
   - (4) → (1): nominative case for the "moved" subject; for other forms of expression in the canonical form, no change.
   - other arguments: forms kept as they are recorded in the canonical frame (except when EFF(jako+4) is used, see above at PAT).

Example of transformation:

požádat (*to ask*)	ACT(1) ADDR(4) PAT(o+4,aby) → ACT(7) ADDR(1) PAT(o+4,aby)
říci (*to say*)	ACT(1) ADDR(3) PAT(o+6) EFF(4,že) → ACT(7) ADDR(3) PAT(o+6) EFF(1,že)
přijímat (*to hire*)	ACT(1) PAT(4) EFF(jako+4) → ACT(7) PAT(1) EFF(jako+1)

Example application to a sentence:

> *Rektor požádal tajemníka o dokumentaci.* → *Tajemník byl požádán rektorem o dokumentaci.* (lit. *The rector asked the secretary for the documentation.* → *The secretary was asked by the rector for the documentation.*)

Here, the ADDR (*the secretary*) moves to the subject position and its form changes from the canonical accusative to nominative (*tajemíka* → *tajemník*), whereas the ACT (*the rector*) must be then expressed in the instrumental case (*rektor* → *rektorem*). The PAT (documentation) remains in the prepositional phrase form, using the preposition *o* (*for*) and the accusative case (o+4).

> *Univerzita přijímala tyto cizince jako překladatele.* → *Tito cizinci byli univerzitou přijímáni jako překladatelé.* (lit. *The university hired these foreigners as translators.* → *These foreigners have been hired by the university as translators.*)

Both the words *cizinci* and *překladatelé* appear in the passive construction in the nominative case, following the ACT(1) PAT(4) EFF(jako+4) → ACT(7) PAT(1) EFF(jako+1) rule.

3.4 Transformation rules for reflexive passivization

In Czech, reflexive passivization adds the particle *se* (lit. *itself*, but it should be noted that this word has lost its proper meaning in this construction) to the verb. It can only be applied to verbs which do not use the same particle in active form (*reflexivum tantum*), or inherent reflexives, where the reflexive meaning is lost completely, such as *smát se*, lit. *to laugh*).

In the transformation of form for reflexive passivization, the subject position is taken by some other argument than ACT, similarly to the periphrastic passivization. However, the ACT never appears on the surface; it is structurally excluded by the syntactic rules of Czech. In other words, this diathesis can be used by the speaker only in the case he does not need to explicitly mention the ACT argument in his utterance, such as in the case when it is general ("everybody").

It is quite common that either PAT or ADDR, which would normally be moved to the subject position, are dropped[7], too. In such a case, a place (LOC, DIR1, ..) or time (TWHEN, TTILL, ..) expression must be present[8], or at least understood from the context. This fact, however, does not change the form of transformation rules.

We do not repeat the individual rules for the transformation of form for the PAT, ADDR, EFF and ORIG arguments, since they are the same as in the periphrastic passivization. However, the following rules must be applied in addition to the periphrastic passivization ones:

1. the particle *se* must be added to the verb (as a separate word), with its word order determined by Czech grammatical rules;
2. the phrase corresponding to ACT must be completely dropped on the surface (i.e., at the tectogrammatical representation the ACT must be represented by the #Gen t-lemma);
3. the tense of the verb remains active (or it remains in the infinitive, as the case may be);
4. the agreement rules on the surface are also determined by the Czech grammatical rules (e.g., if subject is not present at all on the surface, the verb must be in 3rd person singular form, and neuter gender if applicable).

[7] By "dropping" an ACT (PAT, EFF, ...), we mean that there is no word or phrase in the surface form of the sentence corresponding to the dropped argument. In the tectogrammatical representation of the sentence, a special t-lema and a special attribute are used to denote this fact.

[8] Unless the verb itself is in the proper focus of the sentence; for example, as a reply to the question *What do you do with books?*, one can say (only) *Books are read.* without adding when or where.

5. if PAT is moved to the subject position, the same rules apply as in the periphrastic passivization, except the ACT must be dropped on the surface.

Example application to a sentence:

> *Dělníci staví studnu z kamene. → Studna se staví z kamene.* (lit. *The workers build the well from stone. → The well [itself] builds from stone.*)

The ACT (*Dělníci*) must be dropped completely, while the PAT (*Studna*) becomes the subject (in nominative, as usual).

> *Univerzita přijímala tyto cizince jako překladatele. → Tito cizinci se přijímali jako překladatelé.* (lit. *The university hired these foreigners as translators. → These foreigners have [themselves] hired as translators.*)

The ACT (*Univerzita*) has been dropped in the reflexive passivization transformation. In addition and identically to the periphrastic passivization diathesis, both the words *cizinci* and *překladatelé* appear in the passive construction in the nominative case, following the ACT(1) PAT(4) EFF(jako+4) → ACT(7) PAT(1) EFF(jako+1) rule.

> *Děti o Vánocích zpívají koledy. → Koledy se zpívají o Vánocích. Children at Christmastime sing carols. → Carols [themselves] sing at Christmastime.*)

While the PAT undergoes the usual transformation from accusative to nominative, a time (or location) adverbial is usually kept (or "added") for the sentence to sound natural in the reflexive passivization form (for an exception, see the footnote on the preceding page).

3.5 Transformation rules for resultative diathesis

The resultative diathesis (which is normally expressed by the verbs *mít* (lit. *to have*), *dostat* (lit. *to get*) with passive participle of the main verb) is used to move the addressee to the surface subject position (sometimes also to hide the causativity, or the agent, of the event). It can only be used with transitive verbs, where addressee must be present (either as a true addressee (ADDR) with patient (PAT) also present, or shifted to the patient position). Moreover, the *dostat*-type of diathesis can be used for only a limited class of verbs.

The following transformation rules apply (ANY means that the change applies to whatever the original form of the argument was):

1. ADDR is moved to the subject position (no other argument can be moved to this position):
   - (ANY) → (1): nominative case for the "moved" subject

2. PAT, if it appears on the surface, keeps its form, and if in accusative, forces an agreement in gender and case of the passive participle of the main verb, since this argument becomes the complement (Atv, AtvV)[9] on the surface. Moreover, if the gender is feminine in singular, the passive participle must use the special accusative form *-u* (the only verbal form which is thought to have a morphosyntactic features of case). PAT can also be deleted on the surface; the agreement forced on the passive participle form is then neuter singular.
3. ACT is moved to surface object position with the usual form transformation:
   - (1) → (7;od+2)

   In the resultative transformation, however, the (od+2) form is much more frequent than the instrumental case form.
4. Forms of other possible arguments are kept as they are.

Example of transformation (*mít*, lit. *to have*):

připravit (*to prepare*) ACT(1) ADDR(3) PAT(4)
→ ACT(od+2;7) ADDR(1) ?PAT(4)

Example application to a sentence:

Otec připravil dceři školní tašku. → Dcera měla školní tašku připravenu od otce. (lit. *The father prepared [for] daughter the schoolbag. → Daughter had the schoolbag prepared by the father.*)

Example of transformation (*dostat*, lit. *to get*):

přidat (*to add*) ACT(1) PAT(4;na+6) ADDR(3)
→ ACT(od+2;7) ADDR(1) ?PAT(4;na+6)

Example application to a sentence:

Ředitel přidal na platu jen střednímu managementu. → Jen střední management dostal od ředitele přidáno na platu. (lit. *The director raised in salary only to middle management. → Only middle management got from the director raised in salary.*)

3.6 Dispositional diathesis (dispositional modality)

This type of diathesis is used when the speaker expresses the "modality" (extent, in the form of and adverbial) of the relation between the verb and its actor, typically when the actor is an animate object (a human). The adverbial, which must be present on the surface form once the diathesis is applied, expresses often the degree of difficulty (or ease) with which the actor can perform the given action or keep themselves in a given state.

The following transformation rules apply (ANY means that the change applies to whatever the original form of the argument was):

[9] These are the surface syntax functions. For the description of the formalization of surface syntax, which is outside the scope of this article, see e.g. [7].

1. ACT is either elided on the surface (represented as a general actor), or if present, its form is changed to the dative case:
   - (1) → (3): dative case for the actor; often expressed by a pronoun, with no restrictions on the short/long form of the pronoun (standard grammatical rules apply).
2. PAT becomes the surface subject, with accusative becoming nominative:
   - (4) → (1); other forms unchanged.
3. forms for other arguments are unchanged. This might result in two dative-form arguments of the verb, which is normally avoided, but it cannot be excluded on purely grammatical grounds. Often, though, the other arguments are elided on the surface.

Two other conditions must also be fulfilled:

1. An adverbial expressing the degree of difficulty must be present on the surface, or implicitly but clearly understood from the context;
2. the particle *se* must be added to the verb (as a separate word), with its word order determined by Czech syntactic rules. However, if the verb is inherently reflexive (reflexivum tantum), it keeps its single reflexive particle *se* without adding another.

Example of transformation:

studovat (*to study*) ACT(1) PAT(4)
→ (se) ACT(3) PAT(1) adverbial-func(*),

where adverbial-func is typically a modification of manner (MANN).

Example application to a sentence:

Pavel studuje angličtinu. → *Angličtina se Pavlovi studuje snadno.* (lit. *Paul studies English.* → *English (itself) to-Paul studies easily.*)

3.7 Reciprocity

Strictly speaking, reciprocity is not a diathesis in the proper sense, since the changes on the surface are not related to the change in voice or other inflectional feature of the verb. It is considered a diathesis because it shares some of the features of the other types of diatheses (e.g. adding a particle *se* to the verb).

In reciprocity, a single entity (or a group of entities, syntactically expressed as a single phrase, perhaps a coordinated one) occupies two arguments of the verb at the same time: most often it is the actor (ACT) and patient (PAT), but other pairs are also possible, e.g. PAT and ADDR.

Except for adding the particle *se* (unless, similarly to the dispositional modality type, already inherently present), no other changes in form are visible, except that one argument (typically, the “later” one of the two arguments being involved) is missing in the surface and the other one must be either

- in semantic plural (i.e., in surface morphosyntactic plural or it must be a mass noun, which keeps its surface singular inflection);

- a coordination, typically a conjunction, unless all members of the coordination fulfill on of the conditions on this list;
- construction of a single noun phrase modified by another prepositional phrase using the preposition *s(e)* (lit. *with*); this is often considered just another form of coordination;
- a single noun phrase describing a group of people of objects.

In the sentence representation, the "missing" argument is represented by an artificial t-lema `#Rcp` and the appropriate functor (PAT, ADDR, etc.), so that obligatory arguments are present as required. The double-function argument keeps the first functor (ACT in case of ACT–PAT reciprocity, PAT in case of PAT–ADDR reciprocity, etc.) and its form has to fulfill one of the above alternatives. Often, but not necessarily, the adverbials *(sebe/sobě) vzájemně/navzájem* (lit. *(themselves) mutually*), *jeden druhého/druhému/...* (lit. *(to) each other*) or similar are used on the surface to explicitly stress the reciprocity.

Additionally, the forms of location adverbials are slightly changed to reflect the symmetricality of the event, but the description of such changes in free modifiers is beyond the scope of this article (and has not been dealt with systematically in the PDT either).

Example of a transformation:

vidět (*to see*) ACT(1) PAT(4)
→ (se) ACT(<sem-plural>1) PAT(-)

where <sem-plural> is the surface form of all possible semantic plurals, and (-) means there should not be any form corresponding to this argument on the surface (moreover, the t-lema should be `#Rcp`).

Example application to a sentence:

> *Pavel viděl Janu na druhé straně ulice. → Pavel a Jana se viděli přes ulici.* (lit. *Paul saw Jane at the other side of the street. → Paul and Jane saw (themselves) across the street.*)

4 Future Work

In this arcticle, we have described valency in the PDT-Vallex valency dictionary, concentrating on its formal properties and especially on the change of surface form of the verb and its arguments when secondary diatheses are used. The report presented here is not exhaustive; more details can be found in [40].

In the future, we would like to fully formalize the transformation changes not only for the purposes of checking the consistency of annotation (as it was done in the Prague Dependency Treebank and its PDT-Vallex dictionary), but also to serve, in a simpler manner than today, the task of natural language generation, where diathesis and specifically its surface from are of high importance ([31]).

Obviously, we will be extending the PDT-Vallex dictionary further. For example, the dictionary is now being extended to cover all the verbs in the Czech translation of the

Wall Street Journal portion of the Penn Treebank, which is part of a parallel treebank project called the Prague Czech-English Dependency Treebank.[10]. These extensions will also necessarily lead to more examples of diathesis, and thus might need extensions of the surface form transformation rules described in this article.

Acknowledgements

We gratefully acknowledge the support of the Czech Ministry of Education through the grant No. MSM-0021620838, the Grant Agency of Charles University in Prague through the grant No. GAUK 52408/2008 and the Academy of Sciences of the Czech Republic through the grant No. 1ET101120503.

References

[1] Apresjan, J. D. (1995). *Leksičeskaja semantika. Sinonimičeskie sredstva jazyka*. Moskva.

[2] Babby, L. H. (1998). Voice and Diathesis in Slavic. In *Position paper presented at the Workshop on Comparative Slavic Morphosyntax: State of the Art. Indiana University, Spencer*.

[3] Chrakovskij, V. S. (1981). Diateza i referentnost'. (K voprosu o sootnozenii aktivnych, passivnych, refleksivnych i reciproknych konstrukcij). *Zalogovye konstrukcii v raznostrukturnych jazykach*, pages 5–38.

[4] Daneš, F. (1971). Větné členy obligatorní, potenciální a fakultativní (Obligatory, Potential and Optional Constituents of the Sentence). *Miscellanea Linguistica. Acta Universitas Palackiana Olomucensis*, pages 131–138.

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

[6] Grepl, M. and Karlík, P. (1983). *Gramatické prostředky hierarchizace syntaktické struktury věty* . Brno.

[7] Hajič, J. (1998). *Building a Syntactically Annotated Corpus: The Prague Dependency Treebank*. Karolinum, Charles University Press, Prague.

[8] Hajič, J. and Honetschläger, V. (2003). Annotation Lexicons: Using the Valency Lexicon for Tectogrammatical Annotation. *Prague Bulletin of Mathematical Linguistics (PBML)*, 79–80:61—86.

[9] Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., Žabokrtský, Z., and Ševčíková-Razímová, M. (2006). *Prague Dependency Treebank 2.0*. Linguistic Data Consortium, Philadelphia, PA, USA.

[10] Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V., and Pajas, P. (2003). PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In Nivre, J. and Hinrichs, E., editors, *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9 of *Mathematical Modeling in Physics, Engineering and Cognitive Sciences*, pages 57—68, Vaxjo, Sweden. Vaxjo University Press.

[10] http://ufal.mff.cuni.cz/pcedt

[11] Hlaváčková, D., Horák, A., and Kadlec, V. (2006). Exploitation of the VerbaLex Verb Valency Lexicon in the Syntactic Analysis of Czech. In *Lecture Notes in Artificial Intelligence, Proceedings of Text, Speech and Dialogue 2006*, volume 4188, pages 79–86, Berlin and Heidelberg. Springer.

[12] Karlík, P., Nekula, M., and Pleskalová, J. (2002). *Encyklopedický slovník češtiny*. Nakladatelství Lidové noviny, Praha.

[13] Karlík, P., Nekula, M., and Rusínová, Z. (2000). *Příruční mluvnice češtiny*. Nakladatelství Lidové noviny, Praha.

[14] Kettnerová, V. and Lopatková, M. (2009). Changes in Valency Structures of Verbs: Grammar vs. Lexicon. In *This Volume, Proceedings of Slovko 2009*. Springer-Verlag Berlin Heidelberg.

[15] Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

[16] Lopatková, M. and Panevová, J. (2005). Valence vybraných sloves pohybu v češtině. In Piper, P., editor, *Proceedings of Matica Srpska za slavistiky*, pages 1–8, Novi Sad, Serbia and Montenegro, Oct. 27-29.

[17] Lopatková, M., Žabokrtský, Z., Benešová, V., Skwarska, K., Bejček, E., Chvátalová, K., Nová, M., and Tichý, M. (2007a). *Valenční slovník českých sloves*. Karolinum, Praha.

[18] Lopatková, M., Žabokrtský, Z., Benešová, V., Skwarska, K., Bejček, E., Chvátalová, K., Nová, M., and Tichý, M. (2007b). *VALLEX 2.5 - Valency Lexicon of Czech Verbs, version 2.5*.

[19] Mel’čuk, I. A. (1987). *Dependency Syntax: Theory and Practice*. State University of New York Press, Albany.

[20] Mel’čuk, I. A. (2003). Actants. In *Meaning-text theory 2003. Proceedings of the First International Conference on Meaning-text theory*, pages 111–127.

[21] Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová, V., Lopatková, M., Pajas, P., Panevová, J., Razímová, M., Sgall, P., štěpánek, J., Urešová, Z., Veselá, K., žabokrtský, Z., and Kučová, L. (2005). Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka. Technical Report TR-2005-28, Univerzita Karlova v Praze, MFF, ÚFAL, Prague.

[22] Nasr, A. and Rambow, O. (2004). SuperTagging and Full Parsing . In *Proceedings of Seventh International Workshop on Tree Adjoining Grammar and Related Formalisms* .

[23] Padučeva, E. V. (2000). Verbs implying semantic role of result: correlation between diathesis and aspectual meaning. *Linguistische Arbeitsberichte 75*, pages 125–136.

[24] Padučeva, E. V. (2002). Diateza i diatetičeskij sdvig (diathesis and diathesis shift). *Russian Linguistics*, 26:179–215.

[25] Pala, K. and Ševeček, P. (1997). *Valence českých sloves*, pages 41–54. Brno.

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

[27] Panevová, J. (1975). On verbal frames in functional generative description II. *Prague Bulletin of Mathematical Linguistics*, (23):17–52.

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

[29] Panevová, J. (2002). Sloveso: centrum věty; valence: centrální pojem syntaxe. In neuvedeno, editor, *Aktuálne otázky slovenskej syntaxe*, pages x1—x5.

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

[31] Ptáček, J. and Žabokrtský, Z. (2006). Synthesis of Czech Sentences from Tectogrammatical Trees. In *Lecture Notes in Computer Science, Proceedings of the 9th International Conference, TSD 2006*, number 4188 in Lecture Notes in Computer Science, pages 221–228, Berlin / Heidelberg. Springer-Verlag Berlin Heidelberg.

[32] Pustejovsky, J. (1996). *Generative Lexicon*. MIT Press, Cambridge, Massachussetts.

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

[34] Skoumalová, H. (2001). *Czech Syntactic Lexicon. Ph.D. Thesis*. PhD thesis, Faculty of Philosophy, Prague.

[35] Štícha, F. (1984). *Utváření a hierarchizace struktury větného znaku*. Univerzita Karlova, Praha.

[36] Storrer, A. (1992). Verbvalenz. theoretische und methodische grundlagen ihrer beschreibung in grammatikographie und lexikographie. *Reihe Germanistische Linguistik*, 126:414.

[37] Svozilová, N., Prouzová, H., and Jirsová, A. (1997). *Slovesa pro praxi. Valenční slovník nejčastějších českých sloves*. Academia, Praha.

[38] Tesnière, L. (1959). *Eléments de syntaxe structurale*. Klincksieck, Paris.

[39] Urešová, Z. (2005). *Verbal Valency in the Prague Dependency Treebank from the Annotator's Point of View*, pages 93–112. Veda Bratislava, Slovakia.

[40] Urešová, Z. (in prep.). *Valence sloves v Pražském závislostním korpusu*. PhD thesis, Univerzita Karlova v Praze, MFF, Praha.

[41] Uspenskij, V. A. (1977). K ponjatiju diatezy. *Problemy lingvističeskoj tipologii i struktury jazyka*, pages 65–84.

[42] Žabokrtský, Z. (2005). *Valency lexicon of Czech verbs (PhD thesis)*. PhD thesis, Univerzita Karlova v Praze, MFF, ÚFAL, Prague.