ITAT 2013 Proceedings, CEUR Workshop Proceedings Vol. 1003, pp. 10–17
http://ceur-ws.org/Vol-1003, Series ISSN 1613-0073, © 2013 A. Vernerová, M. Lopatková

Towards Automatic Detection of Applicable Diatheses

Anna Vernerová, and Markéta Lopatková

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Czech Republic
{vernerova, lopatkova}@ufal.mff.cuni.cz

Abstract: The valency behavior (argument structure) of lexical items is so varied that it cannot be described by general rules and must be captured in lexicons separately for each lexical item. For verbs, lexicons typically describe only unmarked usage—the active form—while natural languages allow for certain regular changes in the number, type and/or realization of complementations (e.g. passivization). Thanks to their regularity, such changes may be described in a separate rule component of the lexicon; however, they are typically seen in many but not all verbs and their applicability to a given lexical unit (verb meaning) is not predictable from its valency alone. In this paper, we describe our initial experiments with using a large morphologically annotated corpus of Czech for determining which diatheses are applicable to a given lexical unit.

1 Introduction

Valency refers to the argument structure of lexical units.[1] In the Functional Generative Description (FGD), valency belongs to the so-called *tectogrammatical layer* [16, 20], i.e. the layer of linguistically structured meaning. It is captured by so called valency frames specifying the valency complementations (arguments that are either required or specifically permitted by the given lexical unit). For each valency complementation, both its semantics (in the form of a tectogrammatical functor, which captures a coarse-grained semantic role) and its syntactic/morphological form must be specified.

Example 1[2]

- vyzývat ‘to appeal, to challenge’
  ACT_{Nom} ADDR_{Acc} $\text{PAT}_{\text{k+Dat, na+Acc, aby, ať, že}}$
  vyzvat někoho$_{\text{ADDR.Acc}}$, *aby se uklidnil*$_{\text{PAT.aby-Clause}}$
  ‘to ask somebody to calm down’
  vyzvat někoho$_{\text{ADDR.Acc}}$ *na souboj*$_{\text{PAT.na+Acc}}$
  ‘to challenge somebody to a duel’

- apelovat ‘to appeal’
  ACT_{Nom} $\text{ADDR}_{\text{na+Acc}}$ $\text{PAT}_{\text{aby, ať, že}}$
  apelovat na kolegy$_{\text{ADDR.na+Acc}}$, *aby práci dokončili*$_{\text{PAT.aby-Clause}}$ *včas*
  ‘to appeal to his colleagues to finish the work in time’

- apelovat ‘to put emphasis’
  ACT_{Nom} $\text{PAT}_{\text{na+Acc}}$
  v jeho rodině se stále apeluje na morálku$_{\text{PAT.na+Acc}}$
  ‘in his family emphasis is always put on morality’

The above examples demonstrate how valency behavior varies even among semantically close lexical units (LUs), both when they belong to the same lexeme and when they belong to different lexemes. It must therefore be captured for each lexical unit of a verb separately in the form of a lexical entry listed in the valency lexicon. On the other hand, certain changes in the valency structure are regular and can be described in the form of rules which can be specified in a separate component of the lexicon. Such changes are typically seen in many but not all verbs and their applicability to a given lexical unit is not predictable from its valency frame alone.

A lexical entry does not list all of its possible forms but only one—usually the structure corresponding to the active form of the verb, which is considered to be its unmarked use—and a list of rules for creating other possible structures (the marked uses). This description is both economical (less space is needed for storing the information about all available realizations of the LU) and linguistically adequate (it captures generalizations which would not be obvious if all possible surface forms were listed).

Valency lexicons are created with many applications in mind: they help to maintain consistency of corpus annotation, provide syntactic and morphological information during parsing and natural language generation, and may even prove useful in word sense disambiguation and machine translation; moreover, lexicon data is consulted by linguists during their theoretical research and provides useful information for students of Czech. All of these tasks involve actual occurrences of the valency patterns in the natural language, and so the unmarked structures from the lexicon need to be converted into all structures that may appear in the actual data.

A rule based approach to creating derived valency structures has already been used during the annotation of the

[1] Whereas the term *lexeme* roughly corresponds to a dictionary verb item with all its meanings, by a *lexical unit (LU)* we refer to a verb in a given meaning. See Section 3.1 for more details.

[2] The frames and examples are taken from Vallex 2.6, `http://ufal.mff.cuni.cz/vallex/2.6/data/html`.

Prague Dependency Treebank[3] (PDT) [3, 4]. Frames in the valency lexicon PDT-Vallex describe the unmarked structure but all possible structures may appear in actual treebank data. During consistency controls, general rules were used to generate frames describing the marked valency structures; then it was checked whether any of these marked structures matches the data and the annotation in the treebank. (The derived structures carry information about the required form of the verb, and the number and type of the valency complementations including their functors, obligatoriness and permitted forms.) The rules that were used for the conversions are described in detail in [23].

Because correctness of the underlying PDT data was assumed, the rules were allowed to heavily over-generate. For example, "passive" frames of the verb *mít* 'to have' were generated although, in reality, it does not form passive in Czech. While this is a reasonable strategy for consistency checks of annotated data, other tasks that utilize a valency lexicon would benefit from lists of diatheses applicable to any given lexical unit. Manual annotation provides a number of examples of lexical units occurring in different types of diatheses; however, the size of the tectogrammatically annotated PDT data is too small, so we cannot make any conclusions from the fact that a lexical unit does not occur in a diathesis. Therefore, we are trying to draw evidence from a much larger, automatically morphologically annotated corpus. We have decided to use SYN, a non-referential corpus of 1,300 million automatically morphologically tagged words.

For Czech, [21] used simple heuristics for determining which diatheses are applicable to which lexical units (both kinds of passive for verbs with complementations realized as a prepositionless object, infinitive or dependent clause; only reflexive passive for intransitives and verbs where all complementations are realized as prepositional phrases; no passive for reflexives). For other languages, most authors have only studied the applicability of alternations and diatheses to whole lexemes rather than to individual lexical units [11, 14, 15, 19]. We also draw inspiration from the work on automatic extraction of whole frames from corpora, which has been attempted for several languages including English [10], Czech [18], and Polish [2].

2 Diatheses

Regular changes of the valency structure of a lexical unit, in the English-language literature usually called *alternations*, typically allow the speaker to express the same situational meaning (i.e., propositional content characterized by the set of situational participants) in different ways that result in different perspectives from which the situation is viewed. Alternations have already been studied extensively for several decades [12]. In the description of Czech, we follow the classification given by Kettnerová et al. in [7].

Here we focus on *diatheses*—specific relations stemming from the changes in the linking of situational participants, valency complementations and surface syntactic positions. Diatheses belong to the group of *grammaticalized alternations*: they are realized by the use of specific morphological and/or syntactical means, including the grammatical category of voice of the verb and the surface forms of the complementations. They relate different surface syntactic structures of a single lexical unit of a verb. They also belong to the group of *conversive* alternations: the transformation acts as a permutation on the assignment of valency complementations to surface syntactic positions, typically shifting Actor away from the prominent subject position and filling it with some other complementation.

2.1 Types of grammatical diatheses in Czech

In this section, we summarize the description of Czech diatheses as given by [17] and [6], and comment on some of the issues that need to be solved and decisions that need to be made for their automatic analysis.

The unmarked member of the diathesis. The unmarked usage is described in the lexical entry in the lexicon. The verb appears in an active form or as an infinitive; the complementations are realized in the forms specified for them in the lexicon entry. All complementations specified in the entry as obligatory are present on the tectogrammatical layer, although some of them may be elided in the surface realization of the sentence (if their value is either clear from the context or general); inner participants[4] that are not specified in the lexical entry must not appear as arguments of the verb, but free modifications may.

Diatheses with past participle.

1. passive diathesis (periphrastic passive)

   e.g. *Neustále jsem byl někým vyzýván, abych se legitimoval.* 'All the time - I was - someone$_{\text{Instr}}$ - asked - to show my ID.' – 'I kept being asked to show my ID'

   The form of the verb in this diathesis consists of the past participle of the main verb + the verb *být* 'to be' (in a finite or infinite form). The subject slot of the passive construction either remains empty, or it is filled by a complementation which originally filled an object slot (typically that of an Accusative object, but realization through infinitive, clause, genitive, or phrase *jako+Acc* 'as something' is also possible); if the complementation is expressed as a noun phrase, it is turned into the Nominative case. The Actor (which

[3] See `http://ufal.mff.cuni.cz/pdt2.5/` for information about the current version.

[4] Inner participants are complementations with either of the functors ACTot, PATient, ADDRessee, EFFect or ORIGin.

in the active construction fills the subject slot) may be realized either in the Instrumental case, or as a prepositional phrase *od+Gen* 'by/from+Gen'.

2. resultative diathesis with the auxiliary verb *být* 'to be'

   e.g. *Jídlo je uvařeno.* 'The food is cooked.'

   This form of resultative diathesis differs from the periphrastic passive only in meaning, not in the surface form or structure. In many cases, it is not clear which of the two possible readings the speaker has in mind. For example, the sentence *okno je otevřeno* may be interpreted as a case of the resultative diathesis, describing a state, i.e. 'the window is (already) open', or as a case of the passive diathesis, describing an event, i.e. 'the window is (being) opened'. This ambiguity is called *event–state homonymy* in Czech linguistics. Because it is so common, we assume that the passive diathesis is possible whenever the resultative diathesis is possible and vice versa.

   Moreover, Czech also exhibits a competition between past participles and deverbal adjectives. The kind of deverbal adjectives that we have in mind are formed by adding vowel endings to past participles; both the participle and the adjective can then be used to express the resultative meaning, while only the participle can be used to express the passive meaning. On one hand, this interchangeability of the "short" (participle) and "long" (adjectival) forms is often used as a guideline in determining whether a given sentence should be considered resultative—if the participle can be replaced with the adjective, the resultative interpretation is valid. On the other hand, participle forms are sometimes used in purely adjectival meaning, such as in the sentence *stále ještě nebyl najeden* 'he still was not full', which features the word form *najeden* (past participle of the reflexive verb *najíst se* 'satiate oneself, eat so much that one is full'). If we were to read this as a diathesis, this would have to be a case of periphrastic passive or of the resultative diathesis with auxiliary verb *být* 'to be' formed from the sentence *najedl se* 'he ate to be full'. However, it is not possible that the same complementation would fill the subject position in both the active and the passive/resultative diathesis. We have to read this sentence as a sentence with the adjective *najedený* 'full, satiated'.

3. possessive resultative diathesis

   e.g. *Maminka má jídlo uvařeno.* 'The mother has the food cooked.'

   In this type of construction, auxiliary verb *mít* 'to have' is used together with the past participle of the main verb.

   Note that the conversive aspect is crucial for our theoretical concept of a diathesis. For example, only one of the two possible readings of the example sentence above is considered to be a diathesis:

   Mamince včera jídlo připravila tetička. Maminka má tedy již jídlo uvařeno. 'The aunt has prepared the food for the mother yesterday. Therefore, the mother has the food cooked already.' This case is considered to be a diathesis, because the Actor of the first sentence (the aunt) has moved away from the subject position (and is not expressed in the resultative variant at all).

   Maminka vařila celé dopoledne a nyní již má jídlo uvařeno. 'Mother has been cooking all morning and now she already has the food cooked.' This case is not considered to be a diathesis, because the same complementation is corresponding to the subject in both cases.

   In practice, however, it is often impossible to distinguish between the two readings (Panevová et al. [17] claim that out of 60 cases of a resultative diathesis in the PDT, 23 are ambiguous), and the difference is usually only obvious from the context, so it is inaccessible to the kind of naive, syntax-based automatic methods that we are trying to use. Our automatic method does not differentiate between the two readings.

4. recipient passive diathesis

   e.g. *Dostal jsem zaplaceno (od šéfa).* 'I got paid (by the boss).'

   The most visible characteristics of the recipient diathesis is the auxiliary verb *dostat* 'to get' and the past participle. The original frame must contain a complementation in dative or a benefactor; this complementation becomes the subject of the diathetic construction. The actor is expressed in the Instrumental case, or as a prepositional phrase *od+Gen* 'by'. If there is a semantic patient, it keeps its form and agrees with the participle in gender and case. All these conditions together are fairly specific and therefore allow for a fairly accurate search for corpus concordances.

Diatheses with the reflexive particle *se*.

5. deagentive diathesis (reflexive passive)

   e.g. *Vařilo se tu pro emigranty.* 'Cooked - reflexive - here - for - emigrants.' – 'It was cooked here for emigrants.'

   The only surface marks of this diathesis are a verb in the third person (agreeing with the subject in number and gender, or singular neuter for subjectless sentences) and the free reflexive morpheme *se*. The Actor is not expressed in this kind of construction at all. Rules for forming the deagentive diathesis have almost the same conditions as the rules for forming the

passive diathesis (and both types can be applied to almost any frame), and also the ways in which the Patient, Addressee or Effect are moved into the subject position are the same. However, the sets of verbs that allow the two diatheses are different.

6. dispositional diathesis (mediopassive)

   e.g. *Dobře*$_{\text{adverb}}$ *se (mi*$_{\text{Dat}}$*) tu hrál tenis.*
   'Well - reflexive - to-me - here - played - tennis.' – 'For me, this was a good place to play tennis. I enjoyed playing tennis here.'

   A characteristic feature of the dispositional diathesis is an evaluative element, usually an adverb such as *dobře* 'well', *pomalu* 'slowly'. The verb form is the same as in the deagentive diathesis, i.e. a third person verb agreeing with the subject (or signular neuter in subjectless sentences) + reflexive partice *se*. However, the Actor may be expressed on the surface in the dative case.

   Although people do not have difficulties distinguishing between the deagentive and the dispositional diathesis, the difference is hard to grasp for an automatic procedure when the Actor in the dative is elided from the sentence (in a sample from the corpus SYN2005 cited in [17], the Actor was only expressed in 22 sentences out of 143). For example, the following sentence is deagentive, although its surface structure is similar to the example of a dispositional diathesis given above:
   Odpoledne$_{\text{adverb}}$ *se tu hrál tenis.*
   'In the afternoons - reflexive - here - played - tennis.' – 'In the afternoon, tennis was played here.'

   Moreover, this diathesis is very rare—according to [17], it appears only 8 times in the tectogrammatically annotated part of the PDT. We therefore follow the strategy used by Skoumalová [21, p.47] and assume that any imperfective verb which can form the deagentive diathesis can also form the dispositional diathesis.

Diatheses with the reflexive particle *si*.

7. causative diathesis

   e.g. *Nechal si od Gesy vařit.* 'He let - reflexive - by Gesy - cook.' – 'He let Gesy cook for him.'

   This verbal form roughly corresponds to the English 'have something done'. For lexicographic purposes, we view the causative diathesis as a separate sense of the verbs *nechat/dát* 'let/give'. One reason for this treatment lies in the fact that two Actors appear in the construction - the Actor of the verb *nechat/dát* and the Actor of the dependent infinitive.

3 Treatment of diatheses in Vallex and PDT-Vallex

Our proposal is primarily formulated for the purpose of the description of valency in valency lexions of Czech verbs built within the framework of the Functional Generative Description (FGD). We are working with two lexicons, VALLEX 2.6[5], see [13], and PDT-Vallex 2.0[6], see [22], although this phenomenon is to be solved in any valency lexicon. We work with the common format developed for the two lexicons by Bejček et al. [1].

Both lexicons are divided into two components: a data component and a rule component.

3.1 The data component

The data component consists of word entries corresponding to verb lexemes. *Lexeme* is an abstract twofold data structure which associates lexical form(s) and lexical unit(s). *Lexical forms* are all possible manifestations of a lexeme in an utterance (e.g. a lemma or a group of lemmas,[7] all morphological forms of these lemmas, and their reflexive and irreflexive forms). All lexical forms of a lexeme are represented by its lemma(s).

In the lexicon, each *lexical unit* (a sense of a verb) is characterized by a gloss (a verb or a paraphrase roughly synonymous with the given sense) and by example(s) (sentence fragment(s) containing the given verb used in the given sense). The core information on valency characteristics of a lexical unit is encoded as (exactly one) valency frame reflecting the unmarked (active) use of the verb.[8]

In an ideal model of the lexicon, information on the possible application of diatheses is stored in each lexical unit in a special attribute `-diat`. This attribute has not been implemented in either of the lexicons yet, but the attribute `-rfl` (reflexivity) that is present in Vallex overlaps with the proposed attribute `-diat` to a certain extent. It lists possible syntactic functions of the relexive morpheme *se/si*. The values of the `-rfl` attribute are `pass` for reflexive passives in verbs with accusative complements, `pass0` for reflexive passives in intransitive verbs, and `cor4` and `cor3` for cases where *se* and *si* fill the position of an object in accusative (`cor4`) or in dative (`cor3`), showing that the subject is performing an action on itself. If the verb allows for any of these constructions with *se/si*, the possibility has been exemplified with made up examples (the annotators simply converted the examples given for the active diathesis into a passive/reciprocal construction). Many of these

[5] `http://ufal.mff.cuni.cz/vallex/2.6/doc/home.html`

[6] This is the version that has been published as part of the Prague Czech-English Dependency Treebank 2.0.; it is available from `http://ufal.mff.cuni.cz/pcedt2.0/publications/vallex3.xml` and can be browsed at `http://ufal.mff.cuni.cz/lindat/PDT-Vallex.html`.

[7] Vallex lexemes comprise perfective, imperfective and iterative variants, as well as spelling variants, so that the lexicon covers almost twice as many lemmas as lexemes. On the other hand, there is a one-to-one correspondence between lexemes and lemmas in PDT-Vallex.

[8] See the Introduction for more details about valency frames.

Table 1: Counts

	Vallex	**PDT-Vallex**
lexemes[7]	2726	7103
lexical units (LU)	6451	11932

Lemmas

lemmas (L)	4789	7103
LUs (separated by lemma)	11229	11932
reflexive	1528	1590
	31.9 %	22.4 %
nonreflexive, 0 occurrences	586	783
of past participles (tag „Vs“)	12.2 %	11.0 %
nonreflexive, some	888	758
occurrences of Vs, 1 LU	18.5 %	10.7 %
nonreflexive, 0 occurrences	125	71
in sentences with “se”	2.6 %	1.0 %
tagged as “P7-X4-*”		

examples do not sound natural. We hope that our methods will provide some more natural corpus examples. Moreover, we intend to cover other diatheses that the current annotation does not cover.

See Table 1 for counts of lexemes, lemmas and lexical units in both lexicons. Both lexicons are available in machine-tractable XML format and also as human-friendly web pages.

3.2 The rule component

The proposed rule component of the lexicon consists of a set of formal syntactic rules determining changes in the mapping of valency complementations onto surface syntactic positions. They make it possible to obtain all possible surface syntactic manifestations of lexical units of verbs (i.e., number of complementations, their types and possible morphological forms).

At present, we use transformational rules formulated for the purposes of the description of diatheses in PDT-Vallex, the lexicon of the Prague Dependency Treebank, see [23].

4 Methodology

Due to the size of the lexicon, it is preferable to minimize the necessary manual work involved in augmenting the lexicon with information about applicable diatheses. Moreover, experience suggests that annotators tend to be positively biased towards assuming the applicability of the diatheses. Also, examples given by annotators tend to be contrieved/unnatural. To address these problems we would like to have a semiautomatic method which should, where possible

- automatically decide whether a diathesis is applicable,
- provide natural corpus examples of the diathesis to be included in the lexicon,

and in uncertain cases

- provide corpus evidence on the basis of which the annotators can quickly make the decision.

Below we describe such a method in some detail. The method works by iterating over the frames in three passes. The first pass is a negative pass which filters out lexical units where the diathesis is not applicable due to either grammatical concerns or insufficient corpus evidence. The second pass is a positive pass where lexical units with sufficient evidence for applicability are dealt with. In the final step, corpus evidence is gathered for the remaining unclear lexical units. This evidence is then presented to the annotator for a manual decision. If the second or third phase yields a large number of examples, the automatic method should also order them so that simple, clear examples come first. The method of ordering corpus examples used by [8] is well-suited for our purposes.

Due to the difficulties in distinguishing some of the diatheses mentioned above, the proposed semi-automatic procedure only strives to identify cases of the following diatheses: periphrastic passive, possessive resultative, recipient, and deagentive (reflexive passive).

4.1 Negative pass — excluding frames

In the negative pass we use various methods for excluding inapplicable diatheses. In some cases, we exclude whole lexemes (reflexive verbs and lexemes for which no corpus evidence suggesting a possibility of the diathesis was found); the rule-based exclusion, on the other hand, may exclude some lexical units of a lexeme while other proceed into the next phase.

Reflexives. We assume that none of the diatheses is applicable to a lexeme with a reflexive lemma. These cases include reflexiva tantum (*bát se* ‘to fear’) and derived reflexives (*šířit se* ‘to spread (itself)’). This assumption covers 1528 out of 4789 lemmas occurring in Vallex 2.6, and 1590 out of 7116 verb lemmas occurring in PDT-Vallex 2.0. It can be seen from Table 1 that so far this is the most effective step in the negative pass.

We are aware of the fact that this assumption is only approximately valid. According to [9, p. 93], derived reflexives do not form passive (neither periphrastic nor reflexive), but some reflexiva tantum do; [21, p. 43] is only aware of two reflexive verbs that form a periphrastic passive, the reflexiva tantum *tázat se* ‘to ask’ and *obávat se* ‘to fear’, and otherwise assumes that reflexive verbs do not form passives. While [23, p. 124] discusses the limited possibility of forming the reflexive passive of reflexiva tantum, she also gives a (made up?) example of a stylistically non-neutral sentence *smálo se, až se plakalo* ‘it was

laughed so much that it was cried' with reflexive passive of reflexiva tantum.

We have found several other cases where reflexive verbs form a diathesis:

1. *To se lehko pamatuje.* 'This is easy to remember. It is easy to remember it.' (derived from *pamatovat si* 'to remember')
   Na to se lehko zvykne. 'This is easy to get used to. It is easy to get used to it.' (derived from *zvyknout si* 'to get used to')
   Na všechno se zvykne. 'Everything gets used to. People get used to everything.' (derived from *zvyknout si* 'to get used to')
   This usage is almost idiomatic; the first two examples are cases of the dispositional diathesis, and the third seems to be derived from it. We expect that further research will show that this type of construction is productive even among reflexive verbs.

2. *Prezident Václav Havel je lidmi*$_{\text{Instr}}$ *nejméně oblíben od té doby, kdy začal prezidentovat.* 'President Václav Havel is by-the-people least liked since he started presidenting.' – 'President Václav Havel's popularity is the least since he became president.' (derived from *oblíbit si* 'get to like')
   The corpus contains many instances of *je oblíben* 'is liked' which can be easily analyzed as cases of the verbo-nominal predicate *být oblíben(ý)* 'to be liked', not as passive. However, this particular sentence also contains the Actor *lidmi* 'by the people' in the Instrumental case, which is typical of a passive construction. One option is to claim that *lidmi* is a valency complementation of the adjective (*oblíben kým* 'to be liked by whom'). The other option is to admit that this is a case of a passive construction, possibly related to the historical existence of the verb *oblíbit* 'get to like' without a reflexive particle (as documented in [5][9]).

3. *Zdálo se, že toto úsilí už už začne nést ovoce, bylo vděčně povšimnuto čtenáři.* 'It seemed that the effort will soon bear fruit, it was noticed by the readers.' (derived from *povšimnout si* 'notice')
   Here, the reading as a verbo-nominal predicate seems even less likely than in the previous example.

The possibility to form passives of reflexive verbs is certainly an interesting area for further research.

Rule-based exclusion. Some of the diatheses require a particular grammatical structure to be applicable. It is therefore possible to exclude frames where this structure is absent. Here we rely on [23] where a machine-readable list of the necessary structures for each diathesis is compiled. The effectiveness of this exclusion depends on the type of diathesis. The diatheses that are formed with the past participle can be applied to almost any structure. This step is a little more useful in the the cases where the diathesis is formed using the particle *se*.

Corpus-based exclusion. We start with a very naive implementation of this step, excluding the applicability of the diatheses for whole lexemes. Applicability of the diatheses formed with the past participle may be ruled out if the past participle is not found in the corpus. Similarly, we may exclude the applicability of the reflexive passive whenever the verb does not appear in the same sentence as the particle *se* anywhere in the corpus. Table 1 shows that we need to refine these criteria, especially for the exclusion of the reflexive passive.

The mere presence of the *se* token is not necessarily indicative of the given diathesis. First of all, the *se* need not be a particle at all, e.g. in the sentence *tančil se ženou* 'he danced with (his) wife', the word *se* 'with' is in fact a preposition. (The morphological tagger used to tag the corpus SYN is accurate enough to overcome this ambiguity.) But even as a particle, *se* can be part of a different grammatical structure, e.g. in the sentence *snažil se tančit* 'he tried reflexive to dance' the word *se* belongs to the reflexivum tantum *snažit se* 'to try', not to the verb *tančit* 'to dance'. Limiting the search to segments enclosed by punctuation might exclude some genuine examples of diatheses: *minule se, pokud si pamatuji, tančilo až do rána* 'the last time reflexive, as far as I remember, danced until morning' – 'as far as I remember, the last time dancing continued until morning'; we do not want to take the risk of missing some existing evidence already in this phase. Thus, naive corpus search does not suffice to exclude more than a tiny number of verbs (as can be seen from Table 1): auxiliary methods such as (shallow) parsing or at least clause detection are needed. The Prague Dependency Treebank is too small for the purpose of rejecting the applicability of a diathesis, especially if it is rare such as the possessive resultative. (E.g., there are only about 70 instances of the possessive resultative in the whole of PDT.) Corpus SYN, albeit more adequate in size, is not parsed, so a different, inherently less reliable method must be used. We could, for example, base our decision as to whether the *se* is connected to the relevant verb or not on their distance in the sentence.

Combination of the rule-based and corpus-based method. The rules allow us to identify frames describing structures in which a given lexical unit may appear in a diathesis. These structures can be turned automatically into patterns for corpus search. In general, no significant conclusions can be drawn from the fact that the resulting search does not produce any results: Czech is a pro-drop language, so even semantically obligatory elements can

[9] At `http://psjc.ujc.cas.cz/`, search for *oblíbiti* gives 60 instances documented on write-out cards; the relevant entry from the lexicon can be found by searching for *oblíbiti si*.

be elided in the actual sentence. Only the dispositional diathesis contains an element that is obligatory on the surface, but the range of possible morphemic realizations of this evaluative element needs to be further researched.

4.2 Positive copus-based pass

In the positive pass we search the corpus for evidence showing the applicability of a given diathesis. Especially an occurrence of a past participle is indicative of a diathesis (although concerns about the competition between past participles and adjectives need to be addressed). The three kinds of diatheses with past participle forms that we intend to distinguish—periphrastic passive, possessive resultative and recipient passive—moreover differ in the auxiliary verbs. Therefore we assume that instances of past participles found in the corpus can be assigned to a diathesis with a fair amount of certainty. The situation with the passive constructions built with the reflexive particle *se* is more complex, but the techniques developed for the first pass will hopefully help here as well.

The automatic method must be able to assign the evidence found to a particular diathesis and to a particular lexical unit (it does not suffice to know that a verb with many meanings appears in the passive diathesis in the given sentence; we are looking for examples which we can desambiguate). Sometimes, the first pass will give us a single candidate. In other instances, we apply the rules to the remaining frames, derive the description of the full structures corresponding to a diathesis, and then search the corpus for patterns with elements that are unique to only one of the candidates.

4.3 Corpus evidence for manual annotation

Finally, similar methods as in the second phase will be used, but examples with ambiguous status will be output. We expect that the examples will be automatically assigned to a diathesis with high precision. Thus, for each combination of a lexical unit and a diathesis that remain undecided after the previous pass, the system will be able to provide the annotator with a selection of sentences that could be instances of this LU in the given diathesis with high likelihood. The annotator will then either select a couple of examples that demonstrate the applicability of the diathesis, or will decide that the diathesis is not applicable to the given LU.

5 Conclusions

We introduced a (semi-)automatic method for identifying lexical units that undergo individual diatheses, and we have discussed some of the difficulties that stand in the way of a fully automatic procedure. We have also shown that the question whether a diathesis is applicable to a lexical unit may be answered in several different ways:

- The least strict measure is the applicability of a rule for forming the given diathesis. This is a necessary, yet not sufficient condition. The rules have been described in detail in [23] and it is known that they heavily overgenerate.
- If corpus evidence is found for the applicability of the diathesis, the amount/reliability of this evidence may be just as important (especially if the decision is not reviewed by an annotator). Even a single corpus occurrence provides evidence that it is possible to form the diathesis, yet (if the verb itself is frequent) it also provides evidence that for some reason, that possibility is not widely used by the users of the language.
- Lack of corpus evidence leads to the exclusion of some LUs that pass the first test. We expect to find cases where no corpus evidence of the applicability of the diathesis will be found, yet an annotator presented with the LU might still feel that it cannot be excluded completely. (This is essentially the same case as we discussed in the previous paragraph—a possibility that is exploited only rarely—only this time for diathesis-verb combinations that did not appear in the corpus.) We believe that in such a case, and if the entry has been reviewed by an annotator, it is best to provide this information to the user of the lexicon.

Acknowledgments

The research reported in this paper was supported by the grant of the Czech Science Foundation GAČR No. P406/12/0557.

The first author was partially supported by the grant SVV-2013-267314.

This work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013).

References

[1] Bejček, E., Kettnerová, V., and Lopatková, M. (2010). Advanced searching in the valency lexicons using PML-TQ search engine. In Sojka, P., Horák, A., Kopeček, I., and Pala, K., editors, *Text, Speech and Dialogue. 13th International Conference*, volume 6231 of *Lecture Notes in Computer Science*, pages 51–58, Berlin / Heidelberg. Masarykova univerzita, Springer.

[2] Dębowski, Ł. (2009). Valence extraction using EM selection and co-occurrence matrices. *Language resources and evaluation*, 43(4):301–327.

[3] Hajič, J. (2006). Complex corpus annotation: The Prague Dependency Treebank. In Šimková, M., editor, *Insight into Slovak and Czech Corpus Linguistics*, pages 54–73. Veda, Bratislava.

[4] 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. CD-ROM. LDC Catalog No. LDC2006T01.

[5] Hujer, O., Smetánka, E., Weingart, M., Havránek, B., Šmilauer, V., and Získal, A. (1933–1957). *Příruční slovník jazyka českého*. Státní nakladatelství, Státní nakladatelství učebnic, Státní pedagogické nakladatelství, Praha.

[6] Kettnerová, V. and Lopatková, M. (2011). The lexicographic representation of Czech diatheses: Rule based approach. In Majchráková, D. and Garabík, R., editors, *Natural Language Processing, Multilinguality*, pages 89–100, Bratislava, Slovakia. Tribun EU.

[7] Kettnerová, V., Lopatková, M., and Bejček, E. (2012). The syntax-semantics interface of Czech verbs in the valency lexicon. In Fjeld, R. and Torjusen, J., editors, *Proceedings of the 15th EURALEX International Congress*, pages 434–443, Oslo, Norway. Department of Linguistics and Scandinavian Studies, University of Oslo.

[8] Kilgarriff, A., Husak, M., McAdam, K., Rundell, M., and Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In Bernal, E. and DeCesaris, J., editors, *Proceedings of the 13th EURALEX International Congress*, Barcelona, Spain. Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra; Documenta Universitaria.

[9] Kopečný, F. (1962). *Základy české skladby*. Státní pedagogické nakladatelství, Praha, 2. edition.

[10] Korhonen, A. (2002). *Subcategorization Acquisition*. PhD thesis, Ph. D. thesis, University of Cambridge.

[11] Lapata, M. (1999). Acquiring lexical generalizations from corpora: A case study for diathesis alternations. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 397–404. Association for Computational Linguistics.

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

[13] Lopatková, M., Žabokrtský, Z., and Kettnerová, V. (2008). *Valenční slovník českých sloves*. Karolinum, Praha.

[14] McCarthy, D. (2001). *Lexical acquisition at the syntax-semantics interface: diathesis alternations, subcategorization frames and selectional preferences.* PhD thesis, University of Sussex.

[15] McCarthy, D. and Korhonen, A. (1998). Detecting verbal participation in diathesis alternations. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98, pages 1493–1495, Stroudsburg, PA, USA. Association for Computational Linguistics.

[16] Panevová, J. (1994). Valency frames and the meaning of the sentence. In Luelsdorff, P. A., editor, *The Prague School of Structural and Functional Linguistics*, pages 223–243. John Benjamins Publishing Company, Amsterdam, Philadelphia.

[17] Panevová, J. et al. (manuscript). *Syntax současné češtiny (na základě anotovaného korpusu)*. Nakladatelství Karolinum, Praha.

[18] Sarkar, A. and Zeman, D. (2000). Automatic extraction of subcategorization frames for Czech. In *Proceedings of the 18th International Conference on Computational Linguistics, COLING 2000*, volume 2, pages 691–697, Stroudsburg, PA, USA. Association for Computational Linguistics.

[19] Schulte im Walde, S. (2000). Clustering verbs semantically according to their alternation behaviour. In *Proceedings of the 18th conference on Computational linguistics - Volume 2*, COLING '00, pages 747–753, Stroudsburg, PA, USA. Association for Computational Linguistics.

[20] Sgall, P., Bémová, A., Borota, J., Hajičová, E., Hajičová, I., Jirků, P., Panevová, J., Pit'ha, P., Plátek, M., and Vrbová, J. (1986). *Úvod do syntaxe a sémantiky*. Academia.

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

[22] Urešová, Z. (2011a). *Valenční slovník Pražského závislostního korpusu (PDT-Vallex)*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czech Republic.

[23] Urešová, Z. (2011b). *Valence sloves v Pražském závislostním korpusu*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czech Republic.