To Pay or to Get Paid: Enriching a Valency Lexicon with Diatheses

Anna Vernerová, Václava Kettnerová, and Markéta Lopatková

Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské nám. 25, 118 00 Prague 1, Czech Republic
{vernerova, kettnerova, lopatkova}@ufal.mff.cuni.cz

Abstract

Valency lexicons typically describe only unmarked usages of verbs (the active form); however verbs prototypically enter different surface structures. In this paper, we focus on the so-called diatheses, i.e., the relations between different surface syntactic manifestations of verbs that are brought about by changes in the morphological category of voice, e.g., the passive diathesis. The change in voice of a verb is prototypically associated with shifts of some of its valency complementations in the surface structure. These shifts are implied by changes in morphemic forms of the involved valency complementations and are regular enough to be captured by syntactic rules. However, as diatheses are lexically conditioned, their applicability to an individual lexical unit of a verb is not predictable from its valency frame alone. In this work, we propose a representation of this linguistic phenomenon in a valency lexicon of Czech verbs, VALLEX, with the aim to enhance this lexicon with the information on individual types of Czech diatheses. In order to reduce the amount of necessary manual annotation, a semi-automatic method is developed. This method draws evidence from a large morphologically annotated corpus, relying on grammatical constraints on the applicability of individual types of diatheses.

Keywords: diathesis, valency lexicon, syntactic rules

1. Introduction

According to Matthews (2007), *valency* is "the range of syntactic elements either required or specifically permitted by a verb or other lexical unit". The valency behaviour of verbs is so varied that it cannot be described by syntactic rules; on the contrary, it must be captured in valency lexicons separately for each verb. The core information on valency characteristics of a verb can be encoded in the form of valency frames. Although a single lexical unit—a verb in one of its meaning(s)—typically corresponds to a single valency frame, it may appear in different surface structures; this phenomenon is referred to as *alternations*, see esp. the extensive study on English alternations carried out by Levin (1993). In this paper, we concentrate on the alternations of Czech verbs which are expressed by grammatical means, primarily by changes in the grammatical category of verbal voice. We refer to the relationship between the unmarked surface structure (a sentence in the active voice) and the marked surface structure (e.g. a sentence in the passive voice) by the term *diathesis*.

Although diatheses, as (more or less) productive grammatical processes, can be described by explicit syntactic rules, and observations can be made about the syntactic and semantic characteristics of verbs that undergo specific diatheses, their applicability is still often (if not always) lexically conditioned and as such has to be captured in the lexical entries of a lexicon. In this paper, we attempt an explicit description of Czech diatheses in a valency lexicon. The formulation of a formal representation of Czech diatheses will be proposed for the valency lexicon of Czech verbs, VALLEX[1], see esp. (Žabokrtský and Lopatková, 2007). Special attention is devoted to automatically identifying valency frames to which individual types of diatheses are applicable.

After a short overview of existing lexical resources that cover diatheses (Section 1.1.), we briefly describe the VALLEX lexicon (Section 1.2.). The close interplay between grammar and data component of the lexicon in the description of diatheses are demonstrated in (Section 2.) and (Section 3.), respectively. Further, a method of automatic identification of Czech diatheses is demonstrated and its pros and cons are debated in (Section 3.).

1.1. Representation of Diatheses in Lexical Resources

Let us briefly characterize the description of diatheses in the existing lexical resources. The lexicographic representation of these phenomena will be primarily demonstrated on the example of the passive diathesis—being present in typological different languages, it represents a diathesis par excellence.

In many theories, a sharp line is drawn between the lexicon and the grammar. If the passive diathesis is regarded as a regular syntactic transformation, it is not treated in the lexical component, but in the grammar. This approach can be exemplified by the Meaning-Text Theory where passivization (Russian as well as English) is viewed as a productive grammatical process which is not lexically conditioned, and thus should not be treated in the lexicon, an Explanatory Combinatorial Dictionary, but rather in the grammar (Mel'čuk, 2006).

In VerbNet[2] (Kipper et al., 2008), a large database of English verbs which extends the original classification of Levin (1993), the approach to the passive diathesis is very similar to the Meaning-Text Theory (according to Feely et al. (2012), passivization is viewed as a common syntactic transformation not distinctive of verb classes). Yet other

[1] http://ufal.mff.cuni.cz/vallex, http://hdl.handle.net/11858/00-097C-0000-0001-4908-9

[2] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

alternations, e.g. the inchoative alternation, are recorded in the lexical entries for the respective lexical units.

In many projects, the lexical database itself does not contain any specific marking of the possibility to use lexical units in passives (or in other diatheses); however, the passive diathesis is marked in the accompanying corpus annotation. This is the case of PropBank[3] (Palmer et al., 2005), FrameNet[4] (Ruppenhofer et al., 2006) as well as PDT-Vallex[5] (the lexicon linked with the annotation of the Prague Dependency Treebank[6], see (Hajič et al., 2003; Urešová, 2011b)). The information on the possibility to enter the marked member of the passive diathesis is thus only implicitly present.

Let us now introduce the lexical resources that give explicit information on passive constructions.

The Erlangen Valency Patternbank[7] (Herbst and Uhrig, 2009) was extracted from a corpus-based dictionary for language learners. Although only 511 frequent English verbs are covered, the authors still concede that the Cobuild Corpus (which at the time of the compilation of the dictionary contained 320 million words) did not provide sufficient data for the treatment of passive constructions and the lexicographers often had to rely on their intuition. We shall see that data sparsity is a common problem in corpus-based lexicographic treatment of diatheses.

DeepDict[8] (Bick, 2009) is a multilingual co-occurrence lexicon automatically extracted from dependency parsed corpora. Only separate argument slots were extracted (rather than combinations of slots forming frames). One of the slots possibly extracted for a verb is the subject of a passive construction. DeepDict is thus one of few resources which explicitly mark the availability of the passive with a given verb. DeepDict is currently available in 12 languages, including English and Czech.

Besides DeepDict and PDT-Vallex, the information on the passive diathesis of Czech verbs is captured in the VerbaLex valency lexicon[9] (Hlaváčková, 2008) where this information is available for whole synsets of verbs if at least one valency frame in the synset has an accusative object.

Moreover, the applicability of diatheses was also determined during the fully automatically development of the Czech Syntactic Lexicon (Skoumalová, 2002); the algorithm is based on grammatical constraints imposed on arguments of verbs similarly as in Phase 1a of our approach (see Section 3.).

We are not aware of any other lexicon of a Slavic language which would treat the passive diathesis (or other types of diatheses) explicitly.

In this work, we attempt an explicit representation of diatheses in order to provide a comprehensive inventory of all possible surface syntactic manifestations of lexical units of Czech verbs. Below we demonstrate that such a representation requires a close interplay of the lexicon and the grammar and that neither of the two alone is sufficient for a thorough description of the phenomenon.

1.2. VALLEX

VALLEX has been built with the theoretical framework of the Functional Generative Description (FGD). In FGD, valency is related to the so called tectogrammatical layer, a layer of linguistically structured meaning, see esp. (Sgall et al., 1986; Panevová, 1994). Key information on valency structure of a verb is encoded in the form of a valency frame—a single frame corresponds to a single lexical unit of a verb. A valency frame is modelled as a sequence of valency slot(s). Each slot stands for one valency complementation; it consists of a functor (a label for a semantico-syntactic relation), a list of morphemic form(s), and information on obligatoriness. The lexicon also provides information on other syntactic and semantic phenomena, e.g., reciprocity, reflexivity, control, and semantic class membership.

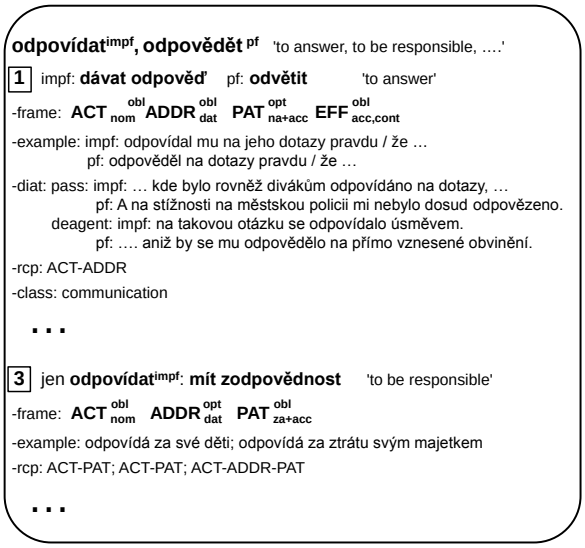

Figure 1: Example of lexical unit in the VALLEX lexicon—verb *odpovídat*^{impf}/*odpovědět*^{pf} 'to answer, to be responsible' (simplified, with proposed attribute `-diat` for diatheses).

The surface syntactic expressions of individual valency slots are implied by morphemic form(s). In the current version of the lexicon, the information on possible morphemic expressions of valency slots describes the uses of the lexical unit of a verb in active voice. However, changes in the category of voice are prototypically associated with changes in valency structure resulting in different surface syntactic structures. Thus a linguistically adequate (and economic) representation of these changes is necessary for a comprehensive description of the valency behaviour of Czech verbs.

Moreover, as the VALLEX lexicon records information on

[3] http://verbs.colorado.edu/~mpalmer/projects/ace.html

[4] https://framenet.icsi.berkeley.edu/fndrupal/

[5] http://hdl.handle.net/11858/00-097C-0000-0023-4338-F

[6] LDC Catalog No. LDC2006T01, http://hdl.handle.net/11858/00-097C-0000-0001-B098-5

[7] http://www.patternbank.uni-erlangen.de

[8] http://gramtrans.com/deepdict

[9] http://nlp.fi.muni.cz/verbalex/htmlDEMO

valency behaviour of 6,460 lexical units represented by 4,520 verb lemmas—covering almost 98% of verb occurrences in the Czech National Corpus—thusthe lexicon enriched with the information on marked structures of diatheses may serve as a basis for NLP tasks such as parsing, machine translation, information extraction and paraphrasing.

2. Diatheses and Their Grammatical Properties: the Grammar Component

In Czech, grammaticalized changes in valency structure of verbs are primarily associated with changes in the morphological category of voice. The relation between surface syntactic structures that differ in the voice category is referred to as diathesis. Diatheses are prototypically associated with a permutation of valency complementations; this permutation affects the prominent syntactic position of subject. We can observe that in case of diatheses, changes in valency structure of verbs are limited only to changes in morphemic forms that imply changes in surface syntactic expressions of the involved valency complementations. In Czech, five types of diatheses are determined according to five marked morphological meanings related to the voice category: passive, resultative, recipient passive, deagentive and dispositional. Syntactic structures with the marked morphological meanings represent the marked members of diatheses, whereas structures with active voice constitute the unmarked members of these relations, see esp. (Panevová et al., 2014).

The marked pairs of the diatheses in the following examples are taken from the Prague Dependency Corpus 3.0[10]:

(1a) `active`
Chalupu v dubnu 1948 v Brně tajně zatkla Státní bezpečnost.
'The State Police arrested Chalupa in Brno secretly in April 1948.'
`passive`
Chalupa byl v dubnu 1948 v Brně tajně
Chalupa was in April 1948 in Brno secretly
zatčen Státní bezpečností.
arrested by State Police
'Chalupa was secretly arrested in Brno in April 1948 by the State Police.'

(1b) `active`
Aby banka poskytla občanovi úvěr, musí (někdo) zajistit jeho návratnost.
'Before a bank grants credit to a customer, someone has to secure its return. '
`resultative`
Aby banka poskytla občanovi úvěr, musí
in order to bank granted citizen credit it must
mít zajištěnu jeho návratnost.
have guaranteed its return
'Before a bank grants credit to a customer, it must have its return guaranteed. '

(1c) `active`
Dále (někdo) lékaři zaplatí za provedené zdravotní zákroky.
'Further, (somebody) will pay the doctor for provided medical interventions. '
`recipient-passive`
Dále dostane lékař zaplaceno za provedené
further will get doctor paid for provided
zdravotní zákroky.
medical interventions
'Further, the doctor will get paid for provided medical interventions '

(1d) `active`
. . . +Sazbu jako takovou by mohli legislativně snížit,
'They could lower the tariff as such in the legislative process,. . . '
`deagentive`
Sazba jako taková by se mohla legislativně
tariff as such COND REFL could legislatively
snížit, . . .
lower . . .
'The tariff as such could be lowered in the legislative process, . . . '

(1e) `active`
Hrál jsem výborně, vůbec se mi nechtělo střídat.
'I was playing perfectly, I didn't want to be substituted at all. '
`dispositional`
Hrálo se mi výborně , vůbec se mi
played REFL to me perfectly , at all REFL to me
nechtělo střídat.
did not feel like substitute
'I was enjoying playing; I did not feel like being substituted at all. '

For the purpose of the representation of diatheses, we propose to divide the lexicon into a data component and a grammar component (Kettnerová et al., 2012). The data component stores information on unmarked members of diatheses, i.e., the uses of lexical units of verbs in active voice. Further, the potential to enter individual diatheses is marked at each lexical unit. The grammar component represents a part of the overall grammar of the language; it contains syntactic rules that determine changes in morphemic forms of valency complementations of verbs associated with diatheses. These rules allow for derivation of valency frames corresponding to marked members of diatheses from the frames describing the unmarked members. As a result, these rules make it possible to obtain all possible surface syntactic manifestations of lexical units of verbs recorded in the lexicon.

[10] `http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3`

2.1. The Recipient Passive diathesis in the Grammar Component

Let us demonstrate the adopted principles on the example of the recipient passive diathesis in (2a) and (2b).

(2a)

Otec	*nařídil*	*Janovi*
ACT(Subj:nom)	PRED(active)	ADDR(Obj:dat)
father	ordered	John

uklidit si pokoj.
PAT(Obj:inf)
to clean up his room

'Father ordered John to clean up his room.'

(2b)

Jan	*dostal*	*od otce*
ADDR(Subj:nom)	PRED(aux)	ACT(Adv:*od*+gen)
John	got	from the father

nařízeno	*uklidit si pokoj.*
PRED(past. part.)	PAT(Obj:inf)
ordered	to clean up his room

'John was ordered by his father to clean up his room.'

Sentences (2a) and (2b) represent surface syntactic structures in the relation of the recipient passive diathesis. Whereas (2a) with active form of the verb *nařídit* 'to order' (in the rule marked as `act`, see Table 1) represents the unmarked member of this relation, (2b) with the recipient passive form of the verb (marked as `rcp-pass`) corresponds to the marked member (formed by the auxiliary verb *dostat*^{pf}*/dostávat*^{impf} 'to get' and the past participle of the main verb). The use of the recipient passive meaning results in a surface syntactic shift of ACTor from the subject position to the less significant adverbial position, while ADDRessee is promoted to the vacated subject. These surface syntactic changes are reflected in changes in morphemic forms of the involved valency complementations: the morphemic expression of ACTor is changed from nominative (2a) into the prepositional group *od*+genitive (2b) and the form of ADDRessee is changed from dative into nominative. On the basis of this observation, the following rule can be formulated:

`Rec-pass d.`	
`verb form`	`replace(act → rcp-pass)`
`ACT`	`replace(nom →` *od*`+gen)`
`ADDR`	`replace(dat → nom)`

Table 1: Rule for the recipient passive diathesis.

This rule allows for generation of the valency frame corresponding to the marked surface structure of the recipient passive diathesis, see (3b) underlying (2b), from the frame representing the unmarked structure, see (3a) describing (2a):

(3a) $\text{ACT}_{\text{nom}}{}^{obl}\ \text{ADDR}_{\text{dat}}{}^{obl}\ \text{PAT}_{\text{acc,inf,dcc}}{}^{obl} \rightarrow$

(3b) $\text{ACT}_{od\text{+gen}}{}^{obl}\ \text{ADDR}_{\text{nom}}{}^{obl}\ \text{PAT}_{\text{acc,inf,dcc}}{}^{obl}$

3. Applicability of Diatheses: the Data Component

Having the lexicon with valency frames describing the unmarked (active) meaning of verbs, a tricky problem arises: which verbs allow for individual marked morphological meanings. We are partly inspired by the method used in the PDT-Vallex lexicon, see esp. (Urešová, 2011a). In this lexicon, grammatical rules and grammatical constraints on their applicability were formulated. However, although grammatical constraints describe necessary conditions for the application of diatheses, they are not sufficient. The rules, which were originally designed for data consistency checking in the Prague Dependency Treebank (PDT), massively over-generate (i.e., they allow also wrong surface structures) and rely on the grammatical correctness of the analysed text. For instance, the recipient passive diathesis is applicable only to the verbs that have either ADDRessee, or PATient expressed in dative in their valency frames. However, many Czech verbs satisfy the given grammatical condition, i.e. they have dative ADDR or PAT in their frames, but they do not form the marked structure of the recipient passive diathesis at all since their semantics does not allow for the recipient passive meaning.

(4a)

Úřad	*odejmul*	*rodině*
ACT(Subj:nom)	PRED(active)	ADDR(InObj:dat)
office	took away	to family

dítě.
PAT(Obj:acc)
child

'The social authorities took the child away from the family.'

(4b) *

Rodina	*dostala*	*odejmuto*
ADDR(Subj:nom)	PRED_1(aux)	PRED_2(past. part.)
family	got	taken away

dítě	*od úřadu.*
PAT(Obj:acc)	ACT(Adv:*od*+gen)
child	by office

'The family got the child taken away by the social authorities.'

Thus although diatheses are productive (or at least semi-productive) grammatical processes that are regular enough to be captured by grammatical rules, they are semantically conditioned: it is semantic properties of a lexical unit that allow an individual type of diathesis to be applied. For this reason, the applicability of a diathesis to lexical units of verbs must be captured in lexical entries in the data component of the lexicon.

Due to the size of the lexicon, it is preferable to minimize the necessary manual work involved in enriching the lexicon with the information on applicable diatheses. Thus we have proposed a semi-automatic method which (i) identifies lexical units for which the applicability of individual diatheses can be excluded without manual intervention, (ii) identifies lexical units that can be supplied with corpus evidence of the diathesis without manual intervention, and (iii) provides corpus evidence for annotators who are asked to decide the remaining cases.

The method proceeds by iterating over the valency frames in three phases.

1. The first phase is a negative phase which filters out lexical units where the processed diathesis is not applicable due to either
   (a) grammatical constraints, or
   (b) insufficient corpus evidence.

   Such lexical units are not further investigated in the remaining two phases.
2. The second phase is a positive phase where lexical units with sufficient evidence for applicability are dealt with.
3. In the final phase, corpus evidence is gathered for the remaining unclear lexical units. This evidence is then presented to annotators for manual decision.

3.1. Applicability of the Recipient Passive Diathesis

Let us demonstrate all three phases on the example of the recipient passive diathesis. This is the rarest of the diatheses that we consider: it has not appeared at all in the part of the Prague Dependency Treebank which has been annotated on the tectogrammatical layer (comprising 833,195 tokens in 49,431 sentences) Therefore, we used for our analyses the largest publicly available synchronic corpus of written Czech: SYN (version 3; 2,685,127,310 tokens), a part of the Czech National Corpus[11].

(ad 1a) The first step of the negative phase is to apply grammatical constraints on lexical units of verbs stored in the VALLEX lexicon. In case of the recipient passive diathesis, two grammatical constraints should be taken into account. The first one is very general as Czech reflexive verbs are considered not to form marked structures of diatheses, with the exception of the dispositional diathesis (Komárek et al., 1986, p. 174 and 177). This constraint cuts down the number of lexical units to which diatheses are applicable to 5009 lexical units out of the overall 6451 units (represented by 3257 verb lemmas out of the overall number of 4781 lemmas).
Second, as stated above, the recipient passive diathesis is applicable only to lexical units that have either ADDRessee, or PATient expressed in dative (see Section 3.). Out of the overall number of 5009 non-reflexive lexical units in VALLEX, only 574 have either dative ADDRessee, or dative PATient in their valency frames; these lexical units are represented by 702 verb lemmas out of the overall number of 3257 non-reflexive lemmas.[12]

(ad 1b) Further, the second step in negative filtering lies in the search of the Czech National Corpus for corpus evidence of the recipient passive diathesis. The result of this search—the set of 432 verb lemmas found in the past participle in distance at most three words to the left or five words to the right of the verb *dostat*^{pf}*/dostávat*^{impf} representing the auxiliary verbs in the recipient passive diathesis—was used as a negative filter of the verb lemmas recorded in VALLEX. Only 270 of these lemmas are covered by VALLEX.

(ad 1) Finally, the intersection of the searches (1a) and (1b) that represents 99 verb lemmas corresponding to 155 lexical units, was selected as the set of candidates for the lexical units to which the recipient passive diathesis is applicable, see Table 2.

(ad 2) In the second (positive) phase, 99 candidate lemmas for the recipient passive diathesis were ranked according to their frequency in the context of the verb *dostat*^{pf}*/dostávat*^{impf} relative to the total frequency in the corpus. For instance, the highest ranking verb *vyčinit* 'to rebuke' exhibits relative frequency 4.76%. On the basis of manual evaluation of 99 searched verb lemmas, the borderline representing a sufficient absolute and relative frequency—at least two occurrences in the context of the verbs *dostat*^{pf}*/dostávat*^{impf} and relative frequency above 0.02%—was established with the aim to achieve as high recall as possible with precision 100%. Only 23 verb lemmas from the overall number of 99 candidate lemmas crossed the stipulated borderline and were accepted as lemmas whose lexical units with dative ADDRessee or PATient allow for the recipient passive diathesis.
Further, corpus occurrences of the 23 selected verb lemmas were added to the lexicon. These occurrences provide corpus evidence of marked structures of the recipient passive diathesis of the lexical units with dative ADDRessee or dative PATient of the given verb lemmas.

(ad 3) In the third phase, corpus evidence for the remaining 76 candidate lemmas with low absolute or relative frequency was manually evaluated. An annotator was asked to indicate whether or not the corpus sentences represent marked structures of the recipient passive diathesis. As a result, we obtain further 43 verb lemmas whose lexical units with dative ADDResse or PATient allow for the recipient passive diathesis.

As a result of the phases (1), (2) and (3) of the experiment, the applicability of the recipient passive meaning was assigned to 66 lemmas in the data component of the VALLEX lexicon. In total, 68 lexical units of these lemmas enter the recipient diathesis. In comparison with the overall number of 574 lexical units with dative ADDRessee or PATient in their valency frames, the overall number of 68 identified lexical units is relatively low.
This experiment has proved that grammatical constraints on the applicability of diatheses are not sufficient and that the possibility to apply a certain type of diathesis on lexical units of verbs should be provided in a lexicon. From a theoretical point of view, it showed that recipient passive structures in Czech, despite being grammaticalized as passive or resultative structures, are rarer than they had been expected in grammar books, see esp. (Daneš et al., 1987, p. 249–251).

[11] https://www.korpus.cz/

[12] In accordance with the Functional Generative Description, pairs or triplets of lemmas that differ only by their aspect (perfective, imperfective and/or iterative) are covered by a single lexeme, which contains one lexical unit for each sense. The corpus-based phases work with lemmas and then map back onto their respective lexical units.

Occurrences in PDT 3.0 t-layer	Recipient diathesis 0		Possessive resultative 82		Passive diathesis 4710	
	Lemmas	Lexical units	Lemmas	Lexical units	Lemmas	Lexical units
in PDT 3.0 t-layer	0	0	50	54	1143	1385
1a VALLEX	702	574	2862	3584	2718	3377
1b SYN	432		3595		12538	
Intersection of 1a and 1b	99	155	1381	2727	2414	3334
2 Candidates with sufficient absolute and relative frequency	23		513		2399	

Table 2: Statistics on the recipient passive diathesis, possessive resultative diathesis and passive diathesis in Czech.

3.2. Applicability of the Possessive Resultative and Passive Diathesis

The same procedure was applied to possessive resultative and passive diathesis.

(ad 1a) In the negative pass through the VALLEX lexicon, we exclude the possibility of the diatheses if given grammatical constraints are not satisfied. In grammars, both possessive resultative and passive diathesis in Czech are associated with transitive verbs whose object should comply with one of the following grammatical conditions, see esp. (Daneš et al., 1987):

- accusative,
- sentential complement with conjunction *aby* or *že*,
- infinitive.

Although the analysis of the corpus data in PDT attests to the grammatical constraints provided by pre-corpus era grammarians, some exceptions are found:

(a) Out of 82 occurrences of the marked members of the possessive resultative diathesis, only 1 does not satisfy the given conditions, see the usage of the intransitive verb *namířit* 'to aim' in (5a),

(b) Out of 4710 occurrences of the marked members of the passive diathesis, 36 lexical units do not have an object in accusative (although some of them have an object expressed in other case than in accusative, see (5b)).

(5a) `resultative diathesis`
namířit ACT(nom) DIR-TO
Po Praze *měl turista namířeno do Bratislavy.*
after Prague had tourist aimed to Bratislava
'After Prague, the tourist was heading to Bratislava.'

(5b) `passive diathesis`
vyhovět ACT(nom) PAT(dat)
Jejich požadavkům bylo vyhověno.
to their requests was met
'Their requests were met.

The number of lexical units and corresponding lemmas in VALLEX that satisfy these conditions can be found in Table 2.

(ad 1b) Similarly as in the case of the recipient passive diathesis, the corpus was searched for concordances containing past participles of verbs in the context of the auxiliary verb, *mít*^{impf}*/mívat*^{iter} 'to have' for the resultative diathesis and *být*^{impf}*/bývat*^{iter} 'to be' for the passive diathesis. The concordances are further filtered to avoid some typical false positives resulting esp. from the fact that the verb *mít* serves also as a modal verb (e.g., *má být uděláno* 'should be done').

(ad 1) The intersection of (ad 1a) and (ad 1a) results in 2727 candidate lexical units corresponding to 1381 verb lemmas for the possessive resultative diathesis and in 3334 candidate lexical units represented by 2414 verb lemmas for the passive diathesis.

(ad 2) Again, the verb lemmas that satisfy the conditions from phase 1 are sorted according to the frequency in which they occur in the specified context relative to their overall frequency in the corpus. A sample of 100 lemmas across the whole range of relative frequencies is manually annotated as either entering or not entering the given diathesis. A cut-off for entering phase 2 is determined as the least relative frequency among these lemmas such that all lemmas with higher relative frequency are annotated as entering the diathesis. The cut-off relative frequency was determined as 0.03% for the resultative diathesis, and 0.02% for the passive diathesis. Lexical units of 513 lemmas are automatically marked as entering the resultative diathesis, and 2399 as entering the passive diathesis, if they satisfy the grammatical conditions as specified in phase 1a.

(ad 3) In contrast to the recipient passive diathesis, the manual evaluation of candidate verb lemmas, i.e., the verb lemmas with low absolute or relative frequency, would be time-consuming with respect to the overall numbers of these candidates. For instance, in case of the possessive resultative diathesis, we obtained 678 of unclear candidate lemmas.

4. Conclusion and Future Work

In this paper, we have proposed a representation of diatheses in the valency lexicon of Czech verbs, VALLEX. We have demonstrated that a close interplay of the data component and the grammar component of the lexicon is necessary. A special attention has been devoted to the applicability of diatheses on Czech verbs. For these purposes,

we have suggested a semi-automatic method of identifying their applicability: the verbs with sufficient corpus evidence that satisfy grammatical conditions have been handled fully automatically whereas the verbs with insufficient corpus evidence are manually evaluated on the basis of automatically selected corpus evidence. When designing this method, the main emphasis has been put on the quality of the resulting lexicon: an attempt is made to determine which cases can be handled automatically with high precision, and the remaining cases are treated manually on the basis of automatically selected corpus evidence

An advantage of our method for determining the lexical units which enter individual diatheses is that it allows us to extract information from corpus data which is only morphologically annotated; thus, much larger sources of data are available than if a syntactically parsed corpus was required. We have demonstrated that although diatheses are grammaticalized in a language, their productivity may oscillate. In general, the more sparse a certain diathesis is, the larger amount of data is needed for determining its applicability. For example, the corpus provides a single concordance witnessing that they enter the recipient diathesis for 14 of the 66 lemmas accepted, and two witnessing concordances for 10 of them. Such rare phenomena are very hard to capture by statistical or machine-learning methods and it is likely that some of this evidence would be lost if we had to rely on more sophisticated NLP tools for its extraction.

The main drawback of the proposed method is that a relatively large amount of manual intervention is required.

Finally, the current representation of diatheses in VALLEX will be enhanced with their relative frequencies as this information is beneficial for both NLP applications (Hajnicz, 2012) and for language learners.

5. Acknowledgements

The research reported in this paper has been supported by the Czech Science Foundation GA ČR, grant No. GA P406/12/0557. The first author has been partially supported by SVV project number 260 104. This work has been using language resources developed and/or stored and/or distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

6. References

Bick, E. (2009). DeepDict—a graphical corpus-based dictionary of word relations. In Jokinen, K. and Bick, E., editors, *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009*, pages 268–271. Northen European Association for Language Technology.

Daneš, F., Hlavsa, Z., and Grepl, M., editors. (1987). *Mluvnice češtiny 3*. Academia, Praha.

Feely, W., Bonial, C., and Palmer, M. (2012). Evaluating the coverage of verbnet. In *Workshop on Interoperable Semantic Annotation*, pages 19–27.

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 *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 University Press.

Hajnicz, E. (2012). Similarity-based method of detecting diathesis alternations in semantic valence dictionary of polish verbs. In Bouvry, P., Kłopotek, M., Leprévost, F., Marciniak, M., Mykowiecka, A., and Rybiński, H., editors, *Security and Intelligent Information Systems*, volume 7053 of *Lecture Notes in Computer Science*, pages 345–358. Springer Berlin Heidelberg.

Herbst, T. and Uhrig, P. (2009). Valency patterns online–the Erlangen Valency Pattern Bank. In *eLEX2009: eLexicography in the 21st century*, page 103, Louvain-la-Neuve. Center for English Corpus Linguistics, Université catholique de Louvain.

Hlaváčková, D. (2008). *Databáze slovesných valenčních rámců VerbaLex*. Ph.D. thesis, Masarykova Universita, Filozofická fakulta, Brno.

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.

Kipper, K., Korhonen, A., Ryant, N., and Palmer, M. (2008). A large-scale classification of english verbs. *Language Resources and Evaluation*, 42(1):21–40.

Komárek, M., Kořenský, J., Petr, J., and Veselková, J., editors. (1986). *Mluvnice češtiny 2*. Academia, Praha.

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

Matthews, P. (2007). *The concise Oxford dictionary of linguistics*. Oxford University Press.

Mel'čuk, I. (2006). Explanatory combinatorial dictionary. *Open problems in linguistics and lexicography*, pages 225–355.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Panevová, J., Hajičová, E., Kettnerová, V., Lopatková, M., Mikulová, M., and Ševčíková, M. (2014). *Mluvnice současné češtiny 2: Syntax na základě anotovaného korpusu*. Karolinum Praha, Prague, Czech Republic.

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

Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C., and Scheffczyk, J. (2006). *FrameNet II: Extended theory and practice (Sept 14, 2010)*. Computer Science Institute, University of California, Berkeley, California.

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

Skoumalová, H. (2002). Verb frames extracted from dictionaries. *Prague Bull. Math. Linguistics*, 77:19–62.

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

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

Žabokrtský, Z. and Lopatková, M. (2007). Valency information in VALLEX 2.0: Logical structure of the lexicon. *The Prague Bulletin of Mathematical Linguistics*, 87:41–60.