Valency Dictionary of Czech Verbs: Complex Tectogrammatical Annotation

Markéta Straňáková-Lopatková, Zdeněk Žabokrtský

Center for Computational Linguistics
Faculty of Mathematics and Physics, Charles University
Malostranské nám. 25, CZ-11800 Prague, Czech Republic
{stranak,zabokrtsky}@ckl.mff cuni.cz
http://ckl.mff.cuni.cz

Abstract

A lexicon containing a certain kind of syntactic information about verbs is one of the crucial prerequisities for most tasks in Natural Language Processing. The goal of the project described in the paper is to create a human- and machine-readable lexicon capturing in detail valency behavior of hundreds most frequent Czech verbs. Manual annotation effort consumed at this project limits the speed of its growth on the one hand, but guarantees significantly higher data consistency than that of automatically acquired lexicons. In this paper, we outline the theoretical background on which the lexicon is based, and describe the annotation schema (lexicon data structure, annotation tools, etc.). Selected quantitative characteristics of the lexicon are presented as well.

1. Introduction

The verb is traditionally considered to be the center of the sentence, and thus the description of syntactic behavior of verbs is a substantial task for linguists. A syntactic lexicon of verbs with the subcategorization information is obviously crucial also for many tasks in Natural Language Processing (NLP) domain. We briefly exemplify the potential contribution of the valency lexicon to several well-known tasks in NLP:

- **Lemmatisation** (choosing the correct lemma for each word in a running text). Example sentences:

  (1) *Stali se matematiky.*
  [They become mathematicians.]

  (2) *Báli se matematiky.*
  [They were afraid of mathematics.]

  In both sentences, the word form *matematiky* occurs. It could be either Acc.pl or Instr.pl of the lemma *matematik* [mathematician] or Gen.sg, Nom.pl, Acc.pl of lemma *matematika* [mathematics]. The lemma can be disambiguated in both sentences using the fact that the verb *stát se* [to become] (sentence 1) contains[1] neither Gen nor Acc in its valency frame, and no frame of the verb *bát se* [to be afraid] (sentence 2) contains Acc or Instr.[2]

- **Tagging** (choosing the correct morphological tag for the given word and lemma). Example:

  (3) *Ptala se jeho bratra.*
  [She asked his brother.]

  The noun phrase *jeho bratra* [his brother] preceded by no preposition can be Gen.sg or Acc.sg. The verb *ptát se* [to ask] allows only the former possibility.

- **Syntactic analysis** (considering a dependency oriented formalism, syntactic analysis can be informally expressed as 'determining which word depends on which'). Examples:

  (4) *Nechala ho spát.*
  'she let him to sleep'
  [She let him sleep.]

  (5) *Začala ho milovat.*
  'she started him to love'
  [She started to love him.]

  In sentence 4 the pronoun *ho* [him] (Gen.sg, Acc.sg) can depend only on the preceding verb *nechat* [to let] (since this verb has a valency frame containing both Acc and infinitive, whereas the valency frame of *spát* [to sleep] contains neither Gen nor Acc). On the other hand, in sentence 5 the same pronoun must depend on the following verb (since no frame of *začít* [to begin] contains both accusative and infinitive). Considering only the morphological tags of the words, both sentences are equivalent. An unambiguous dependency structure[3] cannot be constructed without considering valency frames of the respective verbs.

- **Word sense disambiguation**. Examples:

  (6) *Odpovídal na otázky.*
  [He was answering questions.]

  (7) *Odpovídal za děti.*
  [He was responsible for children.]

  (8) *Odpovídal popisu.*
  [He matched the description.]

[1] In this context, we use 'frame X contains Y' to express the fact that some element of the valency frame X is prototypically realized by the form Y (direct or prepositional case, etc.) on the surface.

[2] The possibility of Nom is excluded in both sentences according to the subject-verb agreement.

[3] A similar claim holds for phrase structure of given sentences.

Different meanings of the same word are often indicated by a change in the valency frames. The meaning of verb *odpovídat* in sentence 6 is 'to answer', in sentence 7 the same word expresses 'to be responsible', and in sentence 8 it expresses 'to match'.

- **'Semantic analysis'**. Examples:

  (9) *Přišel po Petrovi.*
  He came after Peter.

  (10) *Sháněl se po Petrovi.*
  [He seeked for Peter.]

  Prepositional groups most frequently represent adjuncts (as in sentence 9); however, they can also stand for verbal participants (as in 10), which is a crucial difference in most semantically or logically motivated approaches. The role of the prepositional group *po Petrovi* [after / for Peter] cannot be determined without considering valency frames of the respective verbs.

- **Machine translation.** All of the problems mentioned above inevitably arise during any serious attempt at machine translation (MT). Since the existence of a valency dictionary would lead to a higher quality of the respective submodules of such an MT system, it should also increase the quality of the resulting translation.

Existing lexicons for Czech (see Section 4) either do not contain information needed for automatic syntactic analysis, or their coverage is strictly limited, or they are not available in an electronic form, or they are not sufficiently reliable. The consistency is a great problem for most of them.

We present a lexicon of Czech verbs containing rich syntactic information, where the valency information is the most important one. A great emphasis is laid on the formulation of precise criteria for setting the valency frames of particular verbs and their properties, which seems to be a necessary condition for a consistent treatment of the considered phenomena. The lexicon items refer (through Czech WordNet) to EuroWordNet (EWN), which increases the usability of the lexicon for NLP. Emphasis is laid also on both human- and machine-readability of the resulting lexicon.

2. Theoretical Background

2.1. Functional Generative Description

Valency theory is a substantial part of the Functional Generative Description, FGD (Sgall et al., 1986), a dependency oriented description that serves as our theoretical framework. Valency of verbs has been intensively studied since the seventies (Panevová, 1974-75; Panevová, 1980; Panevová, 2001). The concept of valency primarily pertains to the level of underlying representation of a sentence (i.e. the level of linguistic meaning, in FGD called tectogrammatical level). For NLP, also morphemic representation of particular members of the valency frame is important.

The lexical entry for a verb enumerates valency frame(s), at least one but usually more. A valency frame of a verb (in a broader sense) is interpreted as a range of syntactic elements (verbal modifiers) either required or specifically permitted by this verb. It describes a verb in its primary as well as secondary, 'shifted' use (e.g. *tlačit na někoho* [to urge sb / to press on sb]).

The **valency frame** (in a strict sense) of a particular verb consists of valency slots corresponding to inner participants, i.e. actants (both obligatory and optional), and obligatory modifiers (adjuncts, see below).

On the level of underlying representation, we distinguish five **actants** (inner participants) and a wide scale of modifiers. The actants satisfy the following two conditions:

- The combination of actants is characteristic for a particular verb.
- Each actant can appear only once within any occurrence of a particular verb (if coordination and apposition are not taken into account).

The actants distinguished in FGD are Actor (or Actor/Bearer, Act), Patient (Pat), Addressee (Addr), Origin (Orig) and Effect (Eff). Some typical illustrative examples below are taken from the studies of Panevová (quoted in the References).

(11) *Matka*.Act *předělala dětem*.Addr *loutku*.Pat *z Kašpárka*.Orig *na čerta*.Eff.
[Mother.Act re-made a puppet.Pat for children.Addr from a Punch.Orig to a devil.Eff.]

On the contrary, modifiers (e.g. local, temporal, manner, causal) can modify any verb and they can occur repeatedly with the same verb (the constraints are semantically based) - therefore we call them **free modifiers**. Most of them are optional and belong to the 'valency frame' only in a broader sense (for the list of free modifiers see e.g. (Hajičová et al., 2000)). Examples:

(12) *V Praze*.Loc *se sejdeme na Hlavním nádraží*.Loc *u pokladen*.Loc.
[In Prague we will meet at the Main Station near the booking-offices.]

(13) *Kvůli dešti*.Caus *musel čekat pod střechou, protože neměl deštník*.Caus.
'because of rain (he) had to wait under the roof because he didn't have an umbrella'
[As it was raining he had to wait under the roof because he didn't have an umbrella.]

The inner participants can be either **obligatory** (i.e. necessarily present at the level of the underlying representation) or **optional**. Panevová (1974-75) formulated a **dialogue test** as a criterion for the obligatoriness of actants and free modifiers. Informally, the obligatoriness of a modifier means that both the speaker and the listener must know the information expressed by this modifier.[4]

[4] Some of the obligatory participants may be omitted in the surface (morphemic) realization of a sentence, e.g., Actor can be omitted in every Czech sentence. Similarly, free modifiers (both obligatory and optional) are omittable in the surface realization (as e.g. direction for *přijít* [to come], which always means *přijít někam* [to come somewhere]). For the smoothness of the dialogue, both the speaker and the listener must know the necessary information (e.g. from the preceding dialogue or from the broader situation).

	obligatory	optional
inner participants	+	+
free modifiers	+	–

Figure 1: Valency slots creating verbal valency frame (in a strict sense) are marked with '+'

FGD has adopted the concept of **shifting of 'cognitive roles'** in the language patterning (Panevová, 1974-75). Syntactic criteria are used for the identification of Actor and Patient (following the approach of (Tesnière, 1959)), Actor is the first actant, the second is always the Patient. Other inner participants are detected with respect to their semantics

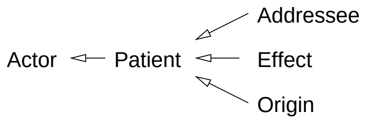

Figure 2: Shifting of cognitive roles.

In other words, if a particular verb has a single actant, it is the Actor (ex. (14)), a verb with two actants has Actor and Patient (regardless the semantics, ex. (15)). The semantics is taken into account with the third and further actants. Examples:

(14) *Škola*.Act *začala.*
[The school lessons began.]

(15) *Bavlně*.Pat *se nic*.Act *nevyrovná.*
[Nothing is as good as cotton.]

(16) *Chlapec*.Act *vyrostl v muže*.Pat
[A boy grew up to a man.]

(17) *Z vašich slov*.Pat *plyne, že zítra nepřijdete*.Act
[It follows from your words that you will not come tomorrow.]

2.2. Enriched Valency Frames

The 'standard' valency view applied in FGD is enriched for the purposes of automatic processing here. In addition to the valency slots creating the valency frame in a strict sense (which does not contain optional free modifiers) also quasi-valency and typical modifiers are stored in the lexicon.

Quasi-valency modifiers are free modifiers that are not obligatory, although they often modify particular verbs and they may specify their meaning (primary, secondary or idiomatic). They can be characterized as 'commonly used modifiers'.

Three sources of quasi-valency modifiers can be distinguished:

- 'usual' modifiers without a strictly specified form (e.g. Direction for verbs of motion, like *jít* [to go]),

- modifiers with a determined morphemic form (e.g. Means in *hrát na kytaru* [play the guitar]), and

- cases with a competition of two occurrences of the modifier, a 'narrower' and a 'wider' specification; the former one is understood as a quasi-valency modifier (e.g. Cause in *zemřít na tuberkulózu kvůli nedostatku léků* [to die of tuberculosis because of the lack of drugs]).

The introduction of **typical modifiers** allows to save all information from the source lexicons. They do not specify the meaning of the verb but they are typical for whole sets of verbs. They usually have a typical form (e.g. Instrumental case for Means as in *psát tužkou* [to write with a pencil], *jet vlakem* [to go by train], or the prepositional group *pro* [for] + Acc for Benefactive as in *pracovat pro firmu* [to work for firm]). In addition, they enable us to capture other syntactic phenomena, such as reciprocity etc. (as described in section 3).

We refer to valency frames capturing valency slots (actants and obligatory free modifiers) as well as quasi-valency and typical modifiers as to **enriched valency frames**.

	obligatory	optional
inner participants	+	+
free modifiers	+	quasi+typical

Figure 3: Modifiers captured in enriched valency frame

For a particular verb, its inner participants have a (usually unique) **morphemic form**, which must be stored in a lexicon (though a prototypical expression of each actant exists, as Nom case for Actor and Acc case for Patient in active sentence, or Dat for Addressee). Free modifiers typically have several different morphemic forms related to the semantics of the modifier. For example, a prepositional group *na* [on] + Acc typically expresses Direction, Prep *v* [in] + Loc has usually local meaning - Where.

The concept of **omissible valency modifiers** is reopened with respect to the task of the lexicon. In principle, conditions of omissibility of particular valency slots on the surface are not yet formally described. We assume that any valency slot is deletable (at least in the specific contexts as e.g. in a question-answer pair).

3. Structure of the Lexicon

3.1. What should a dictionary ideally capture?

The idea is to create lexicon containing all syntactic information useful for NLP. The model proposed offers a complex information on the lexical item (verb), information on its valency frames as well as information specifying elements of these frames.

There is a list of enriched valency frames for each verb (each verb has at least one valency frame, but it may have more frames, with respect to the number of its meanings; primary, secondary as well as idiomatic usage is taken into account).

Several attributes are specified for each valency frame: an ordered sequence of valency slots, a specification of the lexical meaning, examples of usage, the aspectual counterpart, lemma, types of possible diatheses, and pointer(s) to

EuroWordNet synset(s) are the most important ones (see below).

Each frame slot is characterized by a 'functor' (name of an inner participant or modifier, see (Žabokrtsky et al., 2002)), by the type of relation (obligatory, optional and 'quasi-valency' or 'typical' modifier) and by its possible morphemic realization(s).

3.2. Information included in an enriched valency frame

Valency slots. We take over all principles described in section 2. Slots representing valency modifiers are ordered in systemic ordering (introduced in (Sgall et al., 1986)), which reflects unmarked word order in Czech sentence.

Synonyms and examples. A set of synonyms or 'nearly synonyms' together with example(s) of usage specify a particular meaning of the verb.

Alternative frames. A number of verbs exists where a unique meaning can be expressed by two sets of modifiers (e.g. obligatory Addressee and Direction-where often alternates as in *poslal dárky dětem* [he sent gifts to children] / *poslal dárky do Konga* [he sent gifts to Congo]). Such valency frames are marked as alternative frames.

Reciprocity. A concept of reciprocity (Panevová, 1999) expresses the possibility of some modifiers of the given verb to be symmetrical (as in a sentence *Jan a Marie se milují* [John and Mary are in love] where both members *Jan* and *Marie* can be interpreted as Actor and Patient). The possibility of reciprocal use of a verb (in its particular sense) is marked in the lexicon - for relevant valency frames there is a list of modifiers that can be in the relation of reciprocity.

Control. Generally, the notion of control relates to a certain type of predicate (verb of control) and two correferential expressions, a controller and a controllee. We focus on a situation where a verb has an infinitive modifier (regardless its functor). Then controllee is the member that would be the 'subject' of infinitive (which is structurally excluded on the surface), controller is the co-indexed member of the particular valency frame of the head verb (Panevová, 1997); the controller is marked in the lexicon, see also (Skoumalová, 2001). (E.g. the verb *pokoušet se* [to attempt at st] has Patient which can be expressed by an infinitive; its Actor is marked as the controller - see sentence *Marie se pokouší zpívat* [Mary attempts at singing] where *Marie* being the Actor of the head verb *pokoušet se* is the 'subject' of the dependent verb *zpívat* [to sing].)

Diathesis. The lexicon contains valency frames for the active voice of verbs. Many of the diatheses, especially passive constructions are derived regularly (Skoumalová, 2001), thus the individual valency frames are marked only with a marker showing which types of diatheses can be derived from the active form. Only the exceptions are treated explicitly.

Aspectual counterparts. Usually, lexicons designed for human readers list lexical items only for imperfect verbs (which are considered to be the primary ones). The lexicon described here contains separate lexical items for both aspects of verb, the aspectual counterparts are connected with pointers. There are two reasons for this decision:

* **bránit** [to defend / to restrain / to obstruct]
  -aspect: *(imp.)*
+ ACT(1;obl) PAT(4;obl) EFF(před+7,proti+3;obl)
    -synon: *zajišťovat obranu*
    -example: *Obyvatelé brání město před Švédy, před útoky.*
    [The inhabitants defend a town against the Swedes, against attacks.]
    -use: *prim*
    -freq: *3*
    -ewn: *2*
+ ACT(1;obl) ADDR(3;obl) PAT(v+6,Inf,aby;obl) MEANS (7;typ)
    -synon: *zabraňovat, držst zpátky*
    -example: *Brání mu v tom všemi silami.*
    [He impedes him in it with all means.]
    -reciprocity: *ACT-ADDR*
    -control: *ADDR*
    -use: *posun*
    -freq: *15*
    -ewn: *1*
+ ACT(1;obl) PAT(3;obl) MEANS(7;typ)
    -synon: *zabraňovat*
    -example: *Petr bránit jejich štěstí.*
    [Peter obstructs their happiness.]
    -use: *posun*
    -ewn: *1*
* **bránit se** [to prevent]
  -aspect: *(imp.)*
+ ACT(1;obl) PAT(3,proti+3,před+7;opt) MEANS(7;typ)
    -synon: *chránit se*
    -example: *Brání se vydírání; proti vydírání.*
    [They prevent themselves against a blackmail.]
    -use: *prim*
    -freq: *7*

Figure 4: A sample from the valency lexicon

- generally, valency frames may differ for perfect and imperfect aspect of a verb, especially for its secondary or idiomatic usage, and
- the aspectual pairs are treated separately in the Czech WordNet, and thus the pointers to EWN differ for these pairs.

Primary / secondary / idiomatic usage. The valency frames of a particular verb are ordered according to the type of usage - we distinguish primary, secondary and idiomatic usage. This ordering (generally more or less corresponding to the frequency of particular frames - tested on a sample of Czech National Corpus, CNC, (Čermák, 2001)) contribute to an easier orientation in the lexicon. In this stage of work, idiomatic or frozen collocations (where the dependent word is limited either to one lexical unit or to small set of such units, as e.g. *mít na mysli* [to have on mind]) is only partially treated.

Syntactic/semantic classes. Though different semantic classifications of verbs exist, none of them seems to be really appropriate for our task. We preliminarily classify the verbs into several syntactic/semantic classes, such as

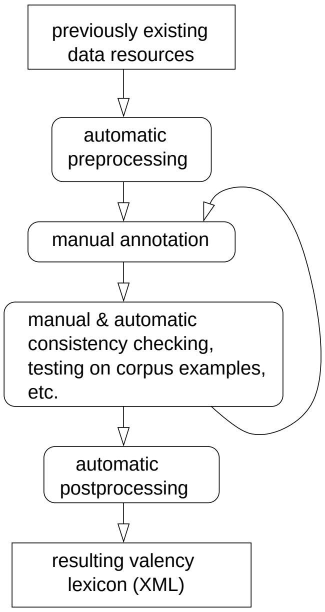

Figure 5: Data flow diagram.

verba dicendi, verbs of movement or verbs of exchange, etc. Such classification helps us when checking the lexicon consistency (verbs from the same class should be treated similarly).

Pointers to Czech WordNet. Valency frames of verbs from the lexicon contained also in Czech WordNet (Pala, Ševeček, 1999) have a pointer to the corresponding Czech synset(s) (=set of synonyms) and through it/them to the interlingual semantic database EuroWordNet (see http://www.hum.uva.nl/ ewn/).

4. How is the Lexicon Created

4.1. Data Resources

Dictionary of verb frames. When creating the lexicon, we utilize other existing electronic resources for Czech. First of all, it is the dictionary of verb frames built up at the Masaryk University (Pala, Ševeček, 1997). The lexicon contains possible morphemic realizations of valency frames of ca 15 000 Czech verbs. Its structure is described in (Horák, 1998). This machine-readable lexicon does not contain information about underlying 'functors' of particular valency frames, the particular meanings of verbs are not specified.[5]

Slovesa pro praxi (Verbs for practise, (Svozilová et al., 1997)). This valency lexicon containing a detailed analysis of ca 750 frequent Czech verbs offers substantial information. Unfortunately, its coverage is limited and the conception of this manually processed lexicon excluded automatic exploitation.

Prague Dependency Treebank. The processing of verbs is based on a number of analyses in theoretical articles concerning FGD, especially those of Panevová. Many unclear aspects are discussed during tectogrammatical annotation of the Prague Dependency Treebank, PDT (Hajičová et al., 2000).

Czech National Corpus. We intensively use the Czech National Corpus, CNC (Čermák, 2001), which serves especially for the verification of valency frames stated and for filling in the gaps.

EuroWordNet and Czech WordNet. The semantic database EuroWordNet (see http://www.hum.uva.nl/ ewn/) and especially its Czech part (Pala, Ševeček, 1999) with its conception of synsets (sets of synonyms, or 'nearly synonyms') contributes to the specification of particular verb meanings.

Slovník české frazeologie a idiomatiky (Lexicon of Czech Phraseology and Idioms, (Čermák, Hronek, 1983)). Though our approach is much more syntax-based, the lexicon of idiomatic expressions helps with the treatment of idioms.

4.2. Annotation

There have been several attempts at creating a valency lexicon automatically but the output of such efforts is not satisfactory. Unfortunately, the great extent of manual annotation seems to be unavoidable for this task, but existing resources can be used which makes it more effective (namely WordNet for Czech, dictionary of morphemic characterization of modifiers of particular verbs, syntactically and morphologically tagged corpora and others).

The lexicon arises in batches of roughly 100 verbs (according to the frequency in the PDT). The 'coverage' of the individual batches is depicted in Figure 6. The process is divided into two steps: automatic preprocessing and manual annotation. In the first step, the resources available are added to all verbs and a preliminary functor assignment is curried on. The second step consists mainly of splitting and merging frames, assigning the functors and correcting the automatically prepared ones, adding the examples. Mapping particular frames on EuroWordNet synset(s) is another important task of the human annotator.

4.3. Software Tools, Data Representation

In order to make the manual annotation as fast as possible, comfortable and effective tools must have been created.

The main annotation tool is the annotation editor. Currently we use a customizable text editor WinEdt (see Figure 7) with a special mode tailored for our lexicon. The data are represented as a (structured) plain text: each line starting with '*' contains a lemma, each line starting with '+' contains a valency frame (written as a sequence of functors followed by parentheses containing surface realization and type of the slot), each line starting with '-' contains a frame attribute (attribute name followed by ':' and attribute value). A (simplified) sample of the data is given in Figure 4.

This approach allows an extremely easy manipulation with lexicon data structures and brings no overhead operations for the annotator. Since the mode colorizes the lexicon data (syntax highlighting), the navigation is also very comfortable.

The second most important tool is the search engine that allows to search for valency frames (in the already ex-

[5]Let us notice also **valency lexicon** that has been **automatically created** on the basis of this dictionary, see (Skoumalová, 2001).

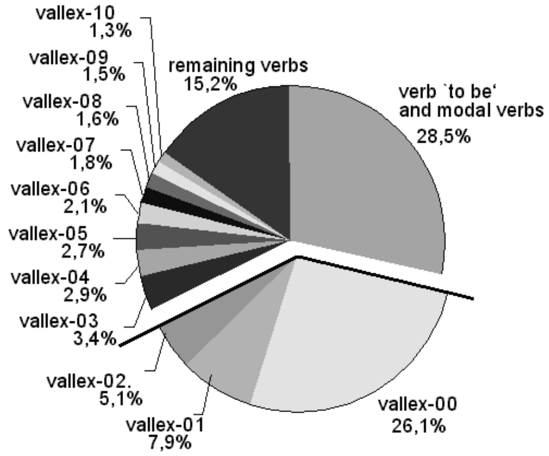

Figure 6: 'Coverage' of the lexicon tested on the verbs in running text from the Czech National Corpus. Vallex-00 contains roughly 160 verbs, each of the remaining batches contains roughly 100 verbs each. The thick line picks out the portion of verbs the annotation of which has been practically finished.

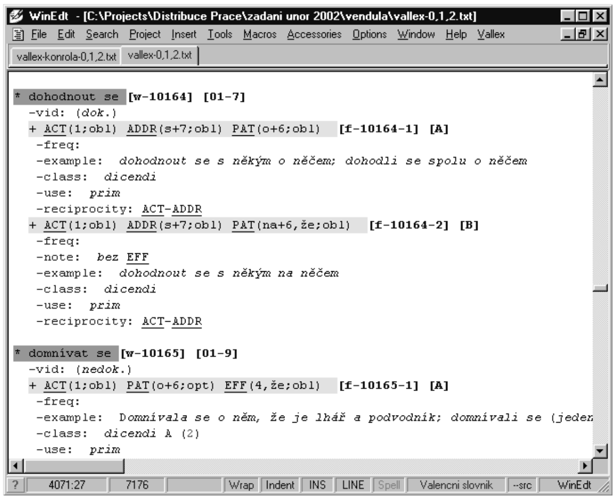

Figure 7: WinEdt screenshot.

isting part of the lexicon) according to a specified query. For example, those frames can be automatically searched which were classified as verba dicendi, have adressee slot expressed by dative.

4.4. Verification, Cross-Checking

We lay a great emphasis on the consistency of the lexicon. The completeness of the data is checked in comparison with the CNC (for each verb a set of sentences is chosen and the annotators 'maps' the occurrences of the verb onto particular valency frames; if need, new frame(s) are added).

The software tools developed allow for sorting valency frames according to a scale of attributes (verb class, morphemic form of modifiers, presence of particular valency slot etc.), which contributes to a consistent treatment of particular phenomena (let us mention e.g. a sometimes unclear boundary between Addressee and Benefactive, or systematic processing of verbs belonging to one class).

The lexicon is used for (manual) tectogrammatical annotation of the PDT. It means a systematic practical verification of the concept accepted as well as of the completeness of the data.

4.5. Selected quantitative characteristics of the data

The project reported on is in progress. The first set of ca 160 verbs served for the development and verification of the annotation scheme, the methodology and the software tools.

At present, a set of 331 most frequent verbs is processed (and used by PDT annotators), as is shown in Figure 6. There are 1110 valency frames for these verbs, which contain altogether 3317 valency slots. Various statistical characteristics are given in Figures 8, 9, and 10.

Another set of 200 verbs is almost completed. Modal verbs and auxiliary *být* [to be], which have been excluded in the first stages as they need a special treatment, is processed now.

We assume that another set of ca 600 verbs will be completed till summer 2002 (it means a 'coverage' of about 85% on the verbs in running text from CNC, see 'remaining verbs' in Figure 6].

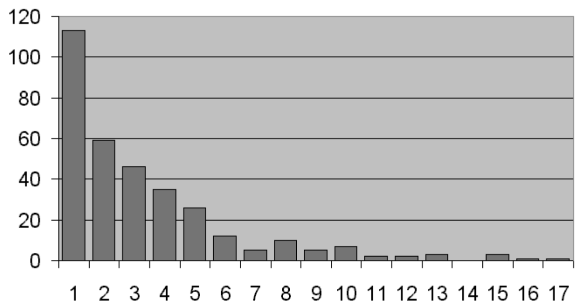

Figure 8: Distribution of the number of valency frames per a lemma.

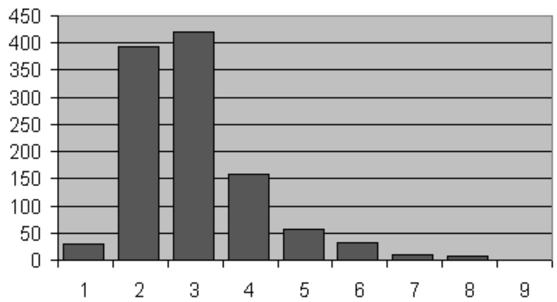

Figure 9: Distribution of the number of valency slots per a frame.

5. Closing remarks

5.1. Open problems

A systematic processing of verbs asks for clear (syntactically based) principles of annotation. Till now, several important questions remain open; though some of them are entirely theoretically described we still miss reliable criteria. The following problems are the most relevant:

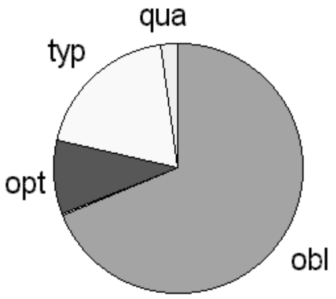

Figure 10: Distribution of values of the type of frame slots.

- The difference between a concrete and an abstract meaning of a verb (e.g. Direction for *vycházet z lesa* [to leave a forest] vs. Direction / Patient for *vycházet z předpokladů* [to start from the premises]).
- Criteria for the distinguishing particular verb meanings (too coarse-grained 'pure syntactic' criteria vs. too fine-grained classification of EWN).
- Criteria for the determination whether a verb with the reflexive particle *se / si* [6] constitutes a separate lexical unit. Example:

  (18) *Matka myje dítě houbou.*
  [Mother washes a child with a sponge.]

  (19) *Myji se každé ráno studenou vodou.*
  [I wash myself every morning with cold water]

  These two Czech sentences exhibit the same syntactic structure; nevertheless, the verbs *mýt* and *mýt se* can be treated in some approaches as two units.
- A complex treatment of idioms.

5.2. Conclusion

We have presented the concept of the lexicon of Czech verbs containing all syntactic phenomena which may be useful for NLP. Though some questions remain open in this stage of our work, the sample of the lexicon (containing 331 most frequent verbs) is successfully used in the process of annotating PDT. A substantial extension is presupposed before summer 2002.

We have mentioned the tasks in NLP to which the lexicon can contribute. On the other hand, it can be useful also for a theoretically based research - the lexicon can be used e.g. for capturing valency of other word classes.

Acknowledgement

The research reported on in this paper has been carried out under the projects MŠMT LN00A063.

6. References

Fr. Čermák. 2001. Language Corpora: The Czech Case. In: V. Matoušek, P. Mautner, R. Mouček, K. Taušer (eds), *Proceedings of the 4th International Conference on Text, Speech and Dialogue - TSD2001*. LNAI 2166. Springer. 21-30.

Fr. Čermák, J. Hronek. 1983. Slovník české frazeologie a idiomatiky (Lexicon of Czech Phraseology and Idioms). Praha. ČSAV.

Fr. Daneš, Z. Hlavsa. 1981. Větné vzorce v češtině (Sentence Patterns in Czech). Praha. Academia.

Ch. J. Fillmore. 1968. The Case for Case. In: E. Bach, R. Harms (editors), *Universals in Linguistic Theory*. New York. 1-90.

E. Hajičová, J. Panevová, P. Sgall. 2000. A Manual for Tectogrammatical Tagging of the Prague Dependency Treebank. UFAL/CKL Technical Report TR-2000-09.

A. Horák. 1998. Verb valency and semantic classification of verbs. In: P. Sojka, V. Matoušek,K. Pala, I. Kopeček, (eds.), *Proceedings of the First Workshop on Text, Speech, Dialogue - TSD'98*. Brno. Masaryk University Press. 61-66.

K. Pala, P. Ševeček. 1997. Valence českých sloves (Valency of Czech verbs). In: *Sborník prací FFBU*. volume A45. Brno. Masaryk University. 41-54.

K. Pala, P. Ševeček. 1999. Final Report Brno, June 1999, Final CD ROM on EWN1,2,LE4-8328. Amsterdam. September 1999.

J. Panevová. 1974-75. On Verbal Frames in Functional Generative Description. Part I. PBML 22. 3-40. Part II. PBML 23. 17-52.

J. Panevová. 1980. Formy a funkce ve stavbě české věty (Forms and Functions in the syntax of Czech sentence). Praha. Academia.

J. Panevová. 1994. Valency Frames and the Meaning of the Sentence. In: P. A. Luelsdorff (ed.), *The Prague School of Structural and Functional Linguistics.* Amsterdam, Philadelphia. Benjamins Publ. Comp. 223-243.

J. Panevová. 1997. More Remarks on Control. In: E. Hajičová, O. Leška, P. Sgall, Z. Skoumalová (eds.), *Prague Linguistic Circle Papers*. Vol. 2. Amsterdam-Philadelphia:John Benjamins. 101-120.

J. Panevová. 1999. Česká recipročná zájmena a slovesná valence (Czech reciprocy pronouns and valency of verbs). Slovo a slovesnost 60. 269-275.

J. Panevová. 2001. Valency Frames: Extension and Reexamination. In: V. S. Charkovskij, M. Grochowski, G. Hentschel (eds.), *Festschrift fuer Andrzej Boguslawski* Studia Slavica Oldenburgensia. No. 9. Oldenburg. Bibliotheks- und Informationssystem. 325-340.

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

H. Skoumalová. 2001. Czech syntactic lexicon. PhD thesis. Prague. Charles University, Faculty of Arts.

H. Skoumalová, Straňáková-Lopatková, Žabokrtský. 2001. Enhancing the Valency Dictionary of Czech Verbs: Tectogrammatical Annotation. In: V. Matoušek, P. Mautner, R. Mouček, K. Taušer (eds), *Proceedings of the 4th International Conference on Text, Speech and Dialogue - TSD2001*. LNAI 2166. Springer. 142-149.

N. Svozilová H. Prouzová, A. Jirsová. 1997. Slovesa pro praxi (Verbs for Practice). Praha. Academia.

[6] Now reflexive passive and reciprocity are not taken into account.

L. Tesnière. 1959. Elements de syntaxe structurale. Paris.

Z. Žabokrtský, P. Sgall, S. Džeroski. 2002. A Machine Learning Approach to Automatic Functor Assignment in the Prague Dependency Treebank. In *Proceedings of LREC 2002*.