Enhancing the Valency Dictionary of Czech Verbs: Tectogrammatical Annotation

Hana Skoumalová[1], Markéta Straňáková-Lopatková[2], and Zdeněk Žabokrtský[2]

[1] Institute of Theoretical and Computational Linguistics, FF UK, Prague
Hana.Skoumalova@ff.cuni.cz
[2] Center for Computational Linguistics, MFF UK, Prague
{stranak,zabokrtsky}@ufal.mff.cuni.cz

Abstract. A syntactic lexicon of verbs with the subcategorization information is crucial for NLP. Two phases of creating such lexicon are presented. The first phase consists of the automatic preprocessing of source data—particular valency frames are proposed. Where it is possible, the functors are assigned, otherwise the set of possible functors is proposed. In the second phase the proposed valency frames are manually refined.

1 Introduction

In this paper[1] we introduce a semi-automatically prepared syntactic lexicon of Czech verbs that is enriched with information about functors (members of valency frames) on the tectogrammatical (underlying) level of language description (Section 2). Such a lexicon is crucial for any applied task requiring automatic processing of natural language. We focus on verbs because of their central role in the sentence—the information about the modifiers of a particular verb enables us to create the 'skeleton' of the analyzed sentence. It can also be used for example in connection with WordNet for semantic grouping of verbs.

As the source data we use a dictionary of verb frames (originally created at Masaryk University) which is automatically preprocessed (Section 3). In the first phase we only process small set of verbs and their frames. This testing set serves for the estimation of the extent of changes in automatically pre-processed valency frames which must be done manually (Section 4). More extensive sets will follow. We expect that a substantially richer lexicon will be available in several months. In the last section (Section 5) the (preliminary) results are presented.

2 The Concept of Valency Frames of Verbs

Valency theory is a substantial part of the Functional Generative Description of Czech (FGD, [Sgall et al, 1986]), and has been intensively studied since the seventies. Originally it was established for verbs and their frames (see esp. [Panevová,

[1] This work has been supported by the Ministry of Education, project LN00A063, and GAČR 405/96/K214.

1974-1975, 1980, 2001]), and was later extended to other parts of speech (nouns and adjectives).

The concept of valency primarily pertains to the level of underlying representation (linguistic meaning) of a sentence and thus it is one of the most important theoretical notions. On the other hand, as the valency information plays a crucial role also for NLP, the morphemic representation of particular members of valency frame is important.

A verbal valency frame (in a strict sense) is formed by so called valency modifiers—that is, the inner participants, either obligatory or optional, together with the obligatory free modifiers. Each Czech verb has at least one valency frame, but it can have more frames. Slots for valency modifiers together with possible morphemic forms of inner participants are stored in a lexicon.

On the level of underlying representation, we distinguish five actants (inner participants) and a great number of free modifiers. The combination of actants is characteristic for a particular verb. Each actant can appear only once in a valency frame (if coordination and apposition are not taken into account). The actants distinguished here are Actor (or Actor/Bearer, Act), Patient (Pat), Addressee (Addr), Origin (Orig) and Effect (Eff). On the contrary, free modifiers (e.g. local, temporal, manner, casual) modify any verb and they can repeat with the same verb (the constraints are semantically based). Most of them are optional and only belong to a 'valency frame' in a broader sense.

The inner participants can be either obligatory (i.e. necessarily presented at the level of underlying representation) or optional. Some of the obligatory participants may be omitted in the surface (morphemic) realization of a sentence if they can be understood as general. Similarly, there exist omissible obligatory free modifiers (as e.g. direction for 'přijít' (to come)). Panevová ([Panevová, 1974-1975]) stated a dialog test as a criterion for the obligatoriness of actants and free modifiers.

FGD has adopted the concept of shifting of 'cognitive roles' in the language patterning ([Panevová, 1974-1975]). 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 semantic roles ([Fillmore, 1968], for Czech [Daneš, Hlavsa, 1981]).

For a particular verb, its inner participants have a (usually unique) morphemic form which must be stored in a lexicon. Free modifiers typically have morphemic forms connected with the semantics of the modifier. For example, a prepositional group Prep 'na' (on) + Accusative case typically expresses Direction, Prep 'v' (in) + Local case has usually local meaning - Where.

In addition to the classical theoretically-based valency also quasi-valency is introduced which may be paraphrased as 'commonly used modification' of a particular item. The concept of quasi-valency enables us to enlarge the information stored in the lexicon, to capture also modifications not belonging to the valency frame in a strict sense ([Straňáková, submitted]). There are free modifiers which are not obligatory (and hence do not belong to the standard valency frame) though they often modify a particular verb. Three sources of such modifiers can be distinguished - (i) 'usual' modifiers without a strictly specified form (like Direction for 'jít' (to go), or Local modifier for 'bydlet' (to stay)), (ii) modifiers with a determined morphemic form (often Regard, e.g., 'zvýhodnit v něčem/na něčem' (to make (st) advantageous for st), or Aim ('potřebovat / poskytovat na něco' (to need / provide (st) for st)), and (iii) theoretically unclear cases with 'wider' and 'narrower' specification (e.g., cause in 'zemřít na tuberkulózu kvůli nedostatku léků' (to die of tuberculosis because of the lack of medicine)).

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)) represent specific phenomenon. We resigned on a very complex task of their processing in this stage.

The concept of omissible valency modifiers is reopened with respect to the task of the lexicon. The omissibility of a modifier is not marked in particular lexical entries—we presuppose that in the surface (morphemic) realization of the sentence any member of valency frame is deletable (at least in the specific contexts as e.g. in a question-answer pair).

Analogically, the fact that particular actant can be realized as a general participant is not marked in the valency frame of a verb.

	obligatory	optional
inner participants	including general participants	+
free modifiers	including omissible modifiers	"commonly used"

Table 1. Verbal modifiers stored in the lexicon.

3 Data Pre-processing

As the source data we use a dictionary of verb frames created at Masaryk University ([Pala and Ševeček, 1997], [Horák, 1998]). The lexicon contains valency frames of circa 15,000 Czech verbs. The structure is described in [Horák, 1998].

3.1 Algorithm for automatic assigning the functors

Identifying and merging frames. In the source lexicon, every lemma is listed only once, even if it has several valency frames. A single valency frame, on the other hand, can have several variants (e.g. 'učit koho co(acc)', 'učit koho čemu(dat)' (to teach sb st)). The variants of one frame are mixed with other frames and thus the first task is to separate the different frames and merge the variants. Let us show it on an example. The verb 'bránit' (to protect/prevent) has the following format in the source lexicon:

```
bránit  <v>hTc3,sI,hPc3-sUeN,hPc3-hTc6r{v},hPTc4,hPTc4-hPTc3r{proti},
hPTc4-hPTc7r{před}
```

Single frames are separated by commas and members inside a single frame are separated by dashes. The attribute 'h' describes 'semantic' features (P-person, T-thing), the attribute ‘c’ stands for morphemic case, ‘r’ means the value of the preposition (in curly braces), ‘sI’ means infinitive and ‘sUeN’ is negative clause with conjunction ‘aby’ (that).

Now, we can arrange the members of all its frames into a table and we can try to find maximal non-intersecting parts.

hTc3			
	sI		
		hPc3 sUeN hPc3 hTc6r {v}	
			hPTc4 hPTc4 hPTc3r{proti} hPTc4 hPTc7r{před}

In the table above we can identify 4 parts. The members that never occur in one frame together can be declared with high probability as variants of one member. Frames with single members (like the first and second frame in the example) can be understood as separate frames, as in the case of ‘mířit kam’ (to head somewhere), ‘mířit na koho’ (to aim at sb), or as variants of one frame, as in the case of ‘bádat nad čím’, ‘bádat o čem’ (to research into st). We decided to ‘merge as much as possible’, because of an easier assignment of the functors. The result is shown below.

```
bránit <v>[hTc3|sI]
bránit <v>[hPc3]-[sUen|hTc6r{v}]
bránit <v>[hPTc4]-[hPTc3r{proti}|hPTc7r{před}]
```

Assigning functors. First, we have to add missing subjects to all frames. Then we assign functors to all members of a frame. Unfortunately, there is no straightforward correspondence between the deep frame and its surface realization, but we can try to find some regularities or tendencies, and then formulate rules for assigning the functors to the surface frames. Among all correspondences between the two levels, there are some which are considered as typical. In the direction from the tectogrammatical level to the morphemic one these are:

Actor → Nominative,
Patient → Accusative,
Addressee → (animate) Dative,
Effect → Prep ‘na’ (to) + Accussative, or Prep ‘v’ (into) + Accusative,
Origin → Prep ‘z’ (from) + Genitive, or Prep ‘od’ (from) + Genitive.

In the opposite direction the correspondences are not so clear because of free modifications, which have a very broad repertory of surface realizations.

For the successful assignment of actants it is necessary to identify free modifiers. The identification is done already during the merging the frames: there exists a list of possible functors for every surface realization, and this list is attached to every member of the original frame. When we merge two members of a frame together we also make an intersection of the attached lists. An empty intersection prevents the two members from being merged. It means that we also get a set of possible functors for every member of a frame as the result of the merging phase. In the optimal case, every member has only one functor assigned.

After identifying free modifiers we can use an algorithm proposed by Panevová and Skoumalová ([Panevová and Skoumalová, 1992]) for the actants. This algorithm is based on the observation that verb frames fall in two categories. The first category contains frames with at most two actants. The functors are assigned on the base of the 'rule of shifting' (see Section 2)—if there is only one actant in the frame it must be an Actor, and if there are two, one of them is an Actor and the other a Patient. As we had to add subjects automatically, we also made the assumption that they all represent Actor, and thus all frames in this category are already resolved.

The other category contains frames with at least three actants, which can be sorted into two subcategories: prototypical and non-prototypical. The prototypical frames contain only typical surface realizations, and the rule about typical realization can be reverted: if the surface frame contains only typical surface forms we can assign the corresponding functors to them. The non-prototypical frames contain at least one untypical surface realization and a different approach must be adopted. The algorithm is described in [Skoumalová, submitted].

After the merging phase, we get three sorts of frames: frames where every member of a frame has only one functor assigned; the second category contains frames with identified actants but ambiguous free modifiers; and the third category contains frames where at least one member is ambiguous between an actant and a free modifier. Approximately one third of all merged frames (circa 6500) falls into the first category ('final' frames in the sequel) and another thousand into the second category. These frames are candidates for further processing with the help of the above mentioned algorithm, and therefore they will be separated from the rest (circa 11,000), which must be left for manual post-editing (the frames belonging to the second and third category are referred as 'ambiguous'). The editor's work should be easier as s/he gets a (small) set of possible functors which can be assigned to every member of a frame and s/he does not have to choose from all 47 possibilities.

3.2 Testing set

For the purpose of testing we made a small set containing 178 most frequent Czech verbs with their frames. We omitted the verb 'být' (to be) as it needs a special treatment, and several modal verbs. The set contained circa 350 frames that were created automatically from the source lexicon. They fall into all three categories mentioned above, which means 1) fully resolved frames, 2) frames with ambiguous free modifiers, and 3) frames with ambiguities between actants and free modifiers.

4 Manual annotation

The data resulting from the preprocessing step are not perfect: they contain incorrectly or ambiguously assigned functors, valency frames proposed may contain mutually excluding (alternating) modifiers, some frames are incorrectly merged into a single one, etc.

That is why we developed a 'tailored' editor for the manual processing of the valency frames of verbs which were pre-processed automatically, as was described above. The editor was implemented as a relational database in Microsoft Access environment.

After obtaining some experiences with annotating the lexicon, we exported the data from the relational database into XML data format (Extensible Markup Language, [Kosek, 2000]). Presently, the XML data are annotated directly in a text editor.

The following attributes are captured for each frame slot:

- functor;
- surface: morphemic realization (mostly morphemic case of a noun or a prepositional group), or a list of possible realization of the particular modifier; the value can be omitted if no special surface realization is required for the given slot (e.g. directional circumstantionals);
- type: this attribute differentiates between obligatory, optional, and quasi-valency modifiers;
- alternative: modifiers, which are mutually excluding, are marked.

4.1 Examples

The following examples illustrate the automatically assigned functors and the manual refinement of valency frames.

The verb 'existovat' (to exist) only has a valency frame that belongs to the first category (fully resolved frames):

existovat R--1[hPTc1]E[hTc2r{u}|hTc6r{na}|hTc6r{v}]$
translated as Actor (Nom) Loc (u+2/na+6/v+6)
manually added mark for arbitrary morphemic realization of local modifier.

The verb 'působit' (to act/operate/work) has been automatically assigned with three valency frames, two of them (1st,3rd) marked as 'ambiguous', one (2nd) as 'final':

působit1 (to operate on st with st) R--1[hPTc1]2CI[hTc7]2A[hPTc4r{na}]& 'ambig.'
translated as Actor (Nom) amb. (na+Acc) amb. (Ins)
manually changed to Actor (Nom) Patient (na+Acc) Means (Ins),
where Actor and Patient are obligatory, Means is a quasi-valency modifier;

působit2 (to do st to sb) R--1[hPTc1]2[hTc4]3[hPc3]& 'final'
translated as Actor (Nom) Patient (Acc) Addr (Dat)
manually the alternative surface forms for Patient are added -
clause attached with conjunctions 'že' (that) or 'aby' (so that);

působit3 (to work as sb)R--1[hPTc1]2P[sU]2JR[hTc4r{jako}]& 'ambig.'
translated as Actor (Nom) amb. (aby) amb. (jako+Acc)
manually changed to Actor (Nom) Patient (jako+Nom) / Loc
where the modifier attached with the conjunction 'aby' belongs to

the second frame (as an alternative representation of Patient),
here the Patient alternates with the Local modifier.

5 Evaluation of Results, Conclusions

In this stage of work only a small testing set of verbs and their frames has been treated. This set serves for clarifying the way of manual processing ('what' and 'how' we want to catch up). After this small lexicon will be brought to perfection it will be used for further development and testing of automatic precedures. But even on this set of available data some preliminary results can be stated.

It is clear now that even the frames marked as 'final' after the pre-processing must be checked and manually refined—about 35 percent of 'final' valency frames were perfect, i.e. 13 percent from all frames proposed. Fortunately, there was a relatively large number of frames which only 'slightly' differ from the issues wanted—approximately 16 percent of valency frames were correctly merged, but the functors were assigned incorrectly (often 'verba dicendi'), in 20 other percent either one functor is missing in the frame, or is superfluous. About 27 percent of frames were deleted (circa one half as incorrect, one half as frames already detected with other morphemic realization). Then the missing frames were manually added and several cycles of corrections followed. We proceeded a cross checking: we extracted and separately compared sets of frames containing a certain functor, we compared frames of verbs with similar meaning etc.

Basic statistical characteristics are presented:

- number of the processed verbs: 178
- number of the frames: 462 (in average 2.6 frames per a verb)
- number of all frame slots: 1481 (in average 3.2 slots per a frame)
- distribution of the number of frame slots per a frame (Table 2)
- distribution of frame slots according to their type (Table 3)
- number of occurences of individual functors in the lexicon (Table 4).

number of slots	1	2	3	4	5	6	7	8
number of frames	16	145	134	95	45	15	10	1
% (out of all frames)	3.5	31.4	29.0	20.6	9.7	3.2	2.2	0.2

Table 2. Distribution of the number of frame slots per a frame.

type	obligatory	optional	quasi-valency
occurences	918	200	363
% (out of all slots)	62.0	13.5	24.5

Table 3. Distribution of the frame slots according to the type.

Roughly one half of the processed verbs is contained in the Czech part of EuroWordNet lexical database [Pala, Ševeček, 1997]. Currently we try to map the valency frames to EuroWordNet synsets.

order	functor	occurences	order	functor	occurences
1	ACT (actor)	460	10	ORIG (origin)	40
2	PAT (patient)	362	11	DIR1 (direction to)	25
3	ADDR (addressee)	93	12	BEN (benefactive)	23
4	EFF (effect)	86	13	AIM (aim)	21
5	MANN (manner)	71	14	ACMP (accompaniment)	18
6	REG (regard)	67	15	TWHEN (time-when)	15
7	LOC (location)	49	16	DIR2 (dir. which way)	14
8	DIR3 (direction from)	49	17	EXT (extent)	13
9	MEANS (means)	48	18	INTT (intention)	7

Table 4. Number of occurences of 18 most frequent functors.

We expect that the large amount of time consumed by the preparation of such a small lexicon has its source in the fact that we have processed the most frequent Czech verbs, which likely belong to the most difficult ones. The extension of data processed may lead (and we hope so) to an increased effectiveness of the algorithm presented.

References

1. Daneš, Fr., Hlavsa, Z.: Větné vzorce v češtině. Academia, Praha, 1981.
2. Fillmore, C.J.: The Case for Case. In: Universals in Linguistic Theory (eds. E. Bach, R. Harms), New York, 1-90, 1968.
3. Horák, A.: Verb valency and semantic classification of verbs. In: TSD'98 Proceedings (eds. Sojka, P., Matoušek, V., Pala, K., and Kopeček, I.), Masaryk University Press, Brno, pp.61-66, 1998.
4. Kosek, J.: XML pro každého. Grada Publishing, Prague, 2000.
5. Pala, K., Ševeček, P.: Valence českých sloves (Valency of Czech verbs). In: Sborník prací FFBU, volume A45, Masaryk University, Brno., pp. 41-54, 1997.
6. Pala, K., Ševeček, P.: Final Report, June 1999, Final CD ROM on EWN1,2,LE4-8328, Amsterdam, September 1999.
7. Panevová, J: On Verbal Frames in Functional Generative Description. Part I, PBML 22, 1974, pp.3-40, Part II, PBML 23, pp. 17-52, 1975.
8. Panevová, J.: Formy a funkce ve stavbě české věty. Academia, Praha, 1980.
9. Panevová, J: Valency Frames: Extension and Re-examination, 2001. (submitted)
10. Panevová, J., Skoumalová, H.: Surface and deep cases. In: Proceedings of COLING '92, Nantes, pp. 885-889, 1992.
11. Sgall, P., Hajičová, E., Panevová, J.: The Meaning of the Sentence in Its Semantic and Pragmatic Aspects (ed. by J. Mey), Dordrecht:Reidel and Prague:Academia, 1986.
12. Skoumalová, H.: Czech syntactic lexicon. PhD thesis, Charles University, Faculty of Arts, Prague. (submitted)
13. Straňáková, M.: Homonymie předložkových skupin a možnost jejího automatického zpracování, PhD thesis, MFF UK (submitted).
14. Tesnière, L.: Élements de syntaxe structurale. Paříž, 1959.
15. Žabokrtský, Z.: Automatic Functor Assignment in the Prague Dependency Treebank. In: TSD2000 Proceedings, LNAI 1906, Springer-Verlag, pp. 45-50, 2000.