Verb Valency Frames in Czech Legal Texts

Eva Mráková and Karel Pala

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
glum@fi.muni.cz, pala@fi.muni.cz
http://www.fi.muni.cz/nlp/

Abstract. This paper deals with valency frames for selected group of Czech verbs belonging to the domain of law. Starting with the lexical database VerbaLex we propose semantic roles for these verbs and formulate their Complex Valency Frames. The lexical database VerbaLex has been developed recently at the NLP Centre FI MU and contains approximately 10 500 Czech verbs. We integrate the proposed 'law' valency frames into it.

Key words: verb valency; Czech language; legal texts

1 Introduction

Though law terms typically consist of the noun and prespositional groups and other nominal constructions, it is necessary to pay attention to the verbs occurring in the legal texts as well. The reason is the following: verbs on one hand do not always display strictly terminological nature, but on the other they are relational elements linking the terminological noun and prepositional groups together. In this respect we can take advantage of the database containing approx. 50 000 law documents and prepared in the Institute of Government and Law Czech Academy of Sciences which we cooperate with within the GACR Grant project PES – GA 407/07/0679.

The verbs from the legal documents were originally processed by the team of F. Cvrček in the Institute of Government and Law. We had received the list of 15 110 items marked as verbs from them. Then we used `ajka` [1] for further processing and obtained the following results: 4 920 items in the list were marked as passive participles - they were not further lemmatized. After manual checking we discovered that 1 611 items from them were not recognized as verbs but for example as adjectives or nouns and they were removed from the list. Thus the list of the correctly recognized verb lemmata finally comprises 10 190 items.

There is a lexical database VerbaLex [2] that has been developed recently in the NLP Centre FI MU and it contains approx. 10 500 Czech verbs with their Complex Valency Frames (CVFs) designed to capture semantic properties of every individual verb. CVFs contain information about morphosyntactic and semantic features of the verb arguments (see below).

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2009, pp. 55–60, 2009.

The idea is to investigate whether and how CVFs designed for 'normal' verbs can be used also for legal verbs and what changes have to be done in semantic labeling of the verb arguments, i. e. what new semantic roles should be added to the ones already utilized in VerbaLex.

2 About VerbaLex

The lexical database VerbaLex consists of the complex valency frames (CVFs) which can be characterized as data structures (tree graphs) describing predicate-argument structure of a verb. They contain the verb itself and its arguments determined by the verb meaning; in Czech their number usually varies from one to five. The argument structure also displays the semantic preferences on the arguments. On the syntactic (surface) level the arguments are most frequently expressed as noun or pronominal phrases in one of the seven cases (in Czech) and also as prepositional cases or adverbials. An example of a complex valency frame for the verb synset *zabít, usmrtit (kill)* capturing both its general and legal meaning looks as follows:

AG<murderer:1>$^{kdo1}_{obl}$ VERB(zabít) PAT<victim:1>$^{koho4}_{obl}$ INS<instrument:1>$^{cim7}_{opt}$
–example: vrah zabil svou oběť nožem (a murderer has killed the victim with a knife).
–synonym: usmrtit
–use: prim.

The semantics of the arguments is typically expressed as belonging to a given semantic role (or deep case), which represents a general role plus subcategorization features (or selectional restrictions). Thus valency frames in VerbaLex include information about:

1. morphosyntactic (surface) information about the syntactic valencies of a verb, i.e. what morphological cases (direct and prepositional ones in highly inflected languages such as Czech) are associated with (required by) a particular verb, and also obligatory adverbials,
2. semantic roles (deep cases) that represent the integration of the general labels with subcategorization features (or selectional restrictions) required by the meaning of the verb.

The inventory of the semantic roles is partly inspired by the Top Ontology and Base Concepts as they have been defined within EuroWordNet project [3]. Thus we work with the general roles like AG, ART(IFACT), SUBS(TANCE), PART, CAUSE, OBJ(ECT) (natural object), INFO(RMATION), FOOD, GARMENT, VEHICLE and others (32). They are combined with the literals from Princeton WordNet 2.0 where literals represent subcategorization features allowing us to climb down the hypero/hyponymical trees to the individual lexical units. For example, we have complex roles like AG(person:1|animal:1) or SUBS(liquid:1) that can be used within the individual CVFs. Their number is approx. 1200.

The verbs in VerbaLex can be characterized as having general, non terminological (legal) meaning. The task then is to take legal verbs and develop the CVFs for them and integrate them into VerbaLex.

3 Some typical legal verbs

We have the following data at our disposal:

- 3,749 verbs occurring only in the legal texts,
- 3,563 verbs occurring only in the VerbaLex,
- 4,830 verbs occurring in the both resources.

We will pay attention only to the first list from which we have chosen a small group of Czech verbs with strictly legal meanings. Two other lists are left aside here. The group of the legal verbs looks as follows (numbers show frequency in legal documents, if we know it):

Table 1. Several examples of legal verbs

verb	frequency	occurs in VerbaLex
žalovat – sue	1064	0
obžalovat – charge	774	0
zažalovat – file a suit, sue	355	0
doznat se – plead guilty	895	0
přiznat se – plead guilty		0
krást – steal		+
okrást – thieve, rob		+
okrádat – steal		+
vykrást – plunder		+
vloupat se – burgle	611	0
vloupávat se		0
znásilnit – rape	585	0
znásilňovat – rape		0
prošetřovat – investigate	306	0
prošetřit – sift, investigate		0
vyšetřit – investigate		+
vyšetřovat – investigate		+
vyslýchat – interrogate		+
odsedět – serve a sentence	231	0
osahávat – grope	159	0
obvinit – accuse		+
obviňovat – accuse		+
vraždit – murder		+
zavraždit – slaughter		+
zabít – kill		+
zabíjet – kill		+

It can be seen that the verbs in the list fall into small subgroups contaning semantically close items – they are either aspect pairs (triples, if iteratives are considered) or prefixed variants. The assumption can be made for them that the individual subgroups will share the complex frames. It also has to be

remarked that for some verbs we know their frequencies only for the perfective or imperfective variant but not for both. The verbs marked with + occur in VerbaLex but only some of their meanings can be considered as legal meanings. This would require more detailed analysis which is a topic for another paper.

There are less frequent verbs in the list of legal verbs that display specialized terminological meanings, for instance the following compound verbs do not occur in the corpus SYN2000 (`http://ucnk.ff.cuni.cz/syn2000.php`) at all: *spoluvinit* (co-accuse), *spoluvázat* (co-bind), *spoluzabezpečovat* (co-ensure), *spoluzaviňovat* (co-cause), *spoluzavazovat* (co-oblige), *spoluzpůsobovat* (co-cause), *spoluzpůsobit* (co-cause, aspect counterpart of the previous one), *spolužalovat* (co-sue), etc. The first member of the given compounds is and adverb *spolu (co-*, thus it will be marked in all CVFs of these verbs.

4 Roles for legal verbs

It can be observed (unpublished report of the F. Cvrček and his team), that the legal verbs co-occur with the nouns, which can be semantically charecterized as follows:

1. one word and multi-word with the autonomous legal meaning, e. g. agreement or contract,
2. nouns with possible legal meaning that follow from the context in which it is used, e. g. person,
3. nouns with clearly non-legal meaning, e. g. chloride,
4. nouns that denote subjects or agents, for instance:
   (a) legal subject such as pachatel (malefactor),
   (b) legal subject following from context, e. g. chairman,
   (c) employment, e. g. sculptor, worker
   (d) legally preferred group, e. g. pensioner,
   (e) subjects by nationality and race, e. g. Serbian, white man,
   (f) nouns with emotional and ideological connotation, such as angel, whore,
   (g) nouns denoting animals, e. g. whale, squirrel, etc.

Some of these characterizations are already included in the VerbaLex and some new should be added (see below).

5 Some CVFs for legal verbs – examples

The above mentioned semantic categories are not too far from the semantic roles as they are used in the CVFs from the VerbaLex database. Thus it can be concluded that the CVFs are suitable for the semantic description of the legal language as well. We can observe the interesting overlaps which allow us to postulate the semantic roles in legal texts. We will mention some of them which can be easily added to the present inventory of the semantic roles in the VerbaLex. This is a positive result, which confirms the assumption that though

legal language displays some specific features it can be analysed with techniques and methods developed for semantic analysis of verb meanings as they occur in a non-terminological use.

In VerbaLex we work with the roles such as AG<person:1> etc., which in legal language correspond to the 'subject' mentioned above. Thus it is possible to take advantage of the roles introduced in VerbaLex and extend them with the semantic categories indicated above. In this way, for instance, we obtain labels such as AG<judge:1>, AG<employee:1> AG<plaintiff:1>, AG<prosecutor:1> AG<defendant:1>, AG<rapist:1>, AG<investigator:3>, AG<thief:1>, AG<policeman:1>, AG<murderer:1> or PAT<victim:1> and similar ones. Then CVFs for the legal verbs mentioned may look as follows (examples):

AG<murderer:1>$^{kdo1}_{obl}$ VERB(zavraždit) PAT<victim:1>$^{koho4}_{obl}$ INS<instrument:1>$^{cim7}_{opt}$,

and similarly one of the possible frames of the verb *obžalovat* (penalize) is

AG<prosecutor:1>$^{kdo1}_{obl}$ VERB(obžalovat) PAT<defendant:1>$^{koho4}_{obl}$ EVENT<crime:1>$^{cim7}_{obl}$.

The mentioned verbs and their CVFs do not occur in the VerbaLex so far. Thus the frames suggested above will be integrated into it. Below, we present two more examples of the CVFs proposed for the specialized legal verbs. It can be seen that the VerbaLex notation is suitable for this purpose:

uložit trest někomu (to condemn somebody to a sentence)

AG<judge:1>$^{kdo1}_{obl}$ VERB PAT<person:1>$^{komu3}_{obl}$ ACT<sentence:1>$^{co4}_{obl}$,

obvinit někoho z trestného činu (to accuse somebody of criminal act)

AG<$^{\texttt{public}}_{\texttt{prosecutor:1}}$>$^{kdo1}_{obl}$ VERB PAT<person:1>$^{koho4}_{obl}$ ACT<act:1>$^{zceho2}_{obl}$.

The described valency frames thus can serve as the descriptions of the meanings of 'legal' verbs – more similar examples can be easily found. For more specialized legal verbs, however, further modifications are needed that require more detailed semantic analysis.

We also have to mention semantic classes [4] of Czech verbs – they represent a sort of 'verbal' ontology. Semantic roles in the valency frames have served as a criterion for finding relevant semantic classes of (Czech) verbs. This can be also applied to law texts in their natural form.

6 Conclusions

To sum up: the goal was to show that valency frames from VerbaLex database can be appropriately applied to the semantic analysis of the legal language. The size (4,830 verbs) of the intersection mentioned above justifies the further investigation. We have already enriched the inventory of the semantic roles in VerbaLex to obtain their more detailed and exact semantic subclassification, or, in other words, their more adequate 'legal' ontology.

Then the roles can be compared with the already existing law ontologies such as the one built within the LOIS (Lexical Ontologies for Legal Information Society) project[1]. In this project, the ontology was built in the WordNet fashion. However, WordNet-like and similar ontologies are structures capturing relations between nouns and noun groups only. We are convinced that more is needed, in particular, a kind of ontology that can be characterized as 'verbal'.

Acknowledgement This research has been funded by the Czech Grant Agency under the Grant Project GA 407/07/0679.

References

1. Sedláček, R.: Morphemic Analyser for Czech. Ph.D. thesis, Faculty of Informatics, Masaryk University, Brno (2005).
2. Horák, A., Hlaváčková, D.: VerbaLex – New Comprehensive Lexicon of Verb Valencies for Czech. In: Computer Treatment of Slavic and East European Languages, Third International Seminar, Bratislava, VEDA (2005) 107–115.
3. Vossen, P., et al.: The EuroWordNet Base Concepts and Top Ontology. Technical Report Deliverable D017, EuroWordNet LE2-4003, University of Amsterdam (1998).
4. Hlaváčková, D., Khokhlova, M., Pala, K.: Semantic Classes of Czech Verbs. In: Proceedings of the IIS Conference 2009, Krakow, in print (2009).

[1] see http://www.ittig.cnr.it/Ricerca/materiali/lois/WhatIsLOIS.htm and also http://nlpweb.kaist.ac.kr/gwc/pdf2006/50.pdf