Recent Advances in Intelligent Information Systems
ISBN 978-83-60434-59-8, pages 41–49

Multilingual Features of Complex Valency Frames

Karel Pala and Aleš Horák

Faculty of Informatics, Masaryk University Brno
Botanická 68a, 602 00 Brno, Czech Republic
{pala,hales}@fi.muni.cz

Abstract

In this paper we deal with the verb valency lexicon of Czech verbs named VerbaLex, which contains complex valency verb frames (CVFs) including both surface and deep valencies.

The most notable features of CVFs include two-level semantic labels with linkage to the Princeton and EuroWordNet Top Ontology hierarchy and the corresponding surface verb frame patterns capturing the morphological cases that are typical of the highly inflected languages like Czech.

We discuss the assumption that CVFs are suitable for a description of the predicate-argument structure not only of Czech verbs but also verbs in other languages, particularly Bulgarian, Romanian and English. We come to the conclusion that this hypothesis can be verified reliably enough exploiting the principle of translatability and also indirectly using semantic classes of (Czech) verbs.

Keywords: verb valency; VerbaLex; complex valency frames

1 Introduction

Semantic role annotation is usually based on the selected inventories of labels for semantic roles (deep cases, arguments of verbs, functors, actants) describing predicate-argument structure of verbs. It can be observed that the different inventories are exploited in different projects, e.g. Vallex (Straňáková-Lopatková and Žabokrtský, 2002), VerbNet (Kipper *et al.*, 2000), FrameNet (Fillmore *et al.*, 2004), Salsa (Boas *et al.*, 2006), CPA (Hanks, 2004), VerbaLex (Hlaváčková and Horák, 2005).

With regard to the various inventories a question has to be asked: how adequately they describe semantics of the empirical lexical data (verbs) as they occur in corpora? It can be seen that some of the inventories should be characterized rather as syntactic than semantic (e.g. Vallex 1.0 or VerbNet). If we are to build verb frames with the goal to describe real semantics of the verbs then we should go 'deeper'. Take, e.g. verbs like *drink* or *eat*, – it is obvious that the role PATIENT that is typically used here labels cognitively different entities – BEVERAGES with *drink* and FOOD with *eat*. If we consider verbs like *see* or *hear* we can observe similar differences. In our view, this situation can be improved if we use more detailed subcategorization features which, however, in lexicons like VerbNet or

Vallex 1.0 are exploited only partially. If we are not able to discriminate the indicated semantic distinctions there can be doubts about the use of the frames with such labels in realistic applications, in other words, the doubts concern descriptive adequacy and expressive power of such notation.

These considerations led us to design the inventory of two-level labels which are presently exploited for annotating semantic roles in Czech verb valency frames in lexical database VerbaLex containing now approx. 10 500 Czech verb lemmata and 19 500 frames.

1.1 Thematic Roles and Semantic types

A question may be asked what is the distinction between “shallow” roles such as AGENT or PATIENT and “deep” roles such as SUBS(food:1), as we use them in VerbaLex. We have already hinted that “shallow” roles seem to be very similar to syntactic functions. At the same time it should be obvious that information that a person functions as an agent who performs an action is both syntactic and partly also semantic. That was the main reason why we included them in our list of the roles. We do not think that SUBS(food:1) is a special case of the deep role, rather, we would like to speak about two-level roles consisting of the ontological part, i.e. SUBS(tantion), and the subcategorization feature part, e.g. beverage:1 which is also a literal in PWN 2.0 that can be reached by traversing the respective hyperonymy/hyponymy tree.

In the Hanks’ and Pustejovsky’s Pattern Dictionary (cf. (Hanks, 2004) and also (Hanks *et al.*, 2007)) a distinction is made between semantic roles and semantic types: “the semantic type is an intrinsic attribute of a noun, while a semantic role has the attribute thrust upon it by the context.” Also lexical sets are distinguished as “clusters of words that activate the same sense of a verb and have something in common semantically.”

Introduction of the mentioned notions is certainly very inspiring in our context, however, we think that at the moment the quoted ’definitions’ as they stand do not seem to be very operational, they are certainly not formal enough for computational purposes. What is needed are the lists of the semantic roles and types but they should be created gradually along with building the necessary ontology/ies. Thus for the time being we have to stick to our two-level roles as they are. They are partly based on the TOP Ontology as used in EuroWordNet project (Vossen, 1998) and partly on the Set of Base Concepts used in EuroWordNet as well.

2 VerbaLex and Complex Valency Frames

The design of VerbaLex verb valency lexicon was driven mainly by the requirement to describe the verb valency frame features in a computer readable form that could be used in the course of automatic syntactic and semantic analysis. After examining actual verb frame repositories for Czech, we have decided to develop *Complex Valency Frames* (CVFs) that contain:

- morphological and syntactic features of the predicate arguments
- two-level semantic labels (roles)

- links to PWN and Czech WordNet hypero/hyponymic (H/H) hierarchy
- differentiation of the animate/inanimate constituents
- default verb position
- verb frames linked to verb senses
- links to the VerbNet classes
- other information mentioned below.

3 Role Annotation and EWN Top Ontology

Presently, our inventory contains about main 40 ontological labels selected from the EuroWordNet Top Ontology (EWN TO), with some modifications, and the 2$^{\text{nd}}$-level subcategorization labels taken mainly from the Set of Base Concepts introduced in EuroWordNet. Their number is approx. 600, they are more concrete and can be viewed as subcategorization features specifying the ontological labels obtained from EWN TO. The motivation for this choice is based on the fact that WordNet has a hierarchical structure which covers approx. 110 000 English lexical units (synsets). It is then possible to use general labels corresponding to the selected top and middle nodes and go down the hyperonymy/hyponymy (H/H) tree until the particular synset is found or matched. This allows us to see what is the semantic structure of the analyzed sentences using their respective valency frames. The nodes that we have to traverse when going down the H/H tree at the same time form a sequence of the semantic features which characterize meaning of the lexical unit fitting into a particular valency frame. In other words, these sequences can be interpreted and used as the special sort of the detailed selectional restrictions.

The ontological labels taken from EWN Top Ontology (about 40) include roles like AGENT, PATIENT, INSTRUMENT, ADDRESSEE, SUBSTANCE, COMMUNICATION, ARTIFACT at the 1$^{\text{st}}$ level. The 2$^{\text{nd}}$-level labels combined with them are represented by complete literals from PWN 2.0 together with their sense number.

An interesting property of the Czech valency frames is that the subcategorization semantic features are endogenous, i.e. they are specified in terms of other synsets of the same WordNet.

The CFVs also contain information about basic metaphors, that are usually characterized as conventional lexical metaphors (Lönneker, 2004). Typically, this includes verbs like *vyletět nahoru (letadlo, náklady), soar (airplane, expenses).* Thus in VerbaLex we have 1817 lemmata and 4681 senses marked with the label "use=fig" denoting lexicalized metaphors.

Other interesting examples of the lexicalized metaphors are *bombardovat hlášeními, informacemi* – in BNC we find *...that the audience be bombarded with information...*, and also cases like *ceny padly – prices dropped* or *ponořit se do emocí – sink in emotions*. Moreover, the notation allows us to capture also metaphors such as *(vláda, škola, banka) budovala ten systém dlouho, (the goverment, school, bank) was building the system for long time* using the role like AG⟨institution:1⟩ or also AG⟨person:1,institution:1⟩.

The frames include the information about idioms as well, their number is presently 1120 (e.g. *klesl na mysli – his spirits sunk*).

The CVF for *drink/pít* then takes the following form:

```
who_nom*AGENT(human:1|animal:1) ⟨drink:1/pít:1⟩
  what_acc*SUBS(beverage:1)
```

4 Can Czech CVFs be Used for Other Languages?

The building of VerbaLex database started during the EU project Balkanet (Balkanet Project, 2002) when about 1500 Czech verb valency frames were included in Czech verb synsets. Starting from Czech WordNet they were linked to English Princeton WordNet and to the WordNets of other languages in Balkanet by means of the Interlingual Index (ILI). We tested a hypothesis that the Czech complex valency frames can be reasonably applied also to the verbs in other languages, particularly to Bulgarian, English and Romanian. Thus, in the Balkanet project an experiment took place in which CVFs developed for Czech verbs have been adopted for the corresponding Bulgarian and Romanian verb synsets (Koeva, 2004) and (Tufis *et al.*, 2006). The results of the experiments were positive (see below the Section 4.1), therefore a conclusion can be made that this can be extended also for other languages.

Experience with Bulgarian and especially Romanian leads us to the view that CVFs developed for Czech can be applied to English equally well. If we exploit ILI and have look at the VFs for Czech/English verbs like *pít/drink, jíst/eat* and apply them to their English translation equivalents we come to the conclusion that the Czech deep valencies describe adequately their meaning as well. VerbaLex is incorporated into Czech WordNet and through ILI also to PWN 2.0, thus we have the necessary translation pairs at hand. This then can also be applied to other WordNets linked to PWN v.2.0 (or higher). Thus we rely on the principle of translatability which here means that for most of the synsets it is possible to find their lexicalized translation equivalent, i.e. the deep valencies developed for the individual Czech verbs can be reasonably exploited also for their English equivalents. There is a problem with surface valencies which in English are based on the fixed order SVOMPT and on morphological cases in Czech but this can be considered rather as a technical issue at the moment.

4.1 Bulgarian example

The enrichment of Bulgarian WordNet with verb valency frames was initiated by the experiments with Czech WordNet (CzWN) which, as we said above, already contained approx. 1500 valency frames (cf. (Koeva *et al.*, June 2004)). Since both languages (Czech and Bulgarian) are Slavonic the assumption was that a relatively large part of the verbs should realize their valency in the same way. The examples of Bulgarian and Czech valency frames in the Figure 1 show that this assumption has been justified (English equivalents come from PWN 1.7). It should be remarked that Bulgarian is in fact caseless but this fact did not play an important role in this experiment.

produce, make, create – create or manufacture a man-made product
BG: {proizveždam} njakoj*AG(person:1)| neščo*ACT(plant:1) =
neščo*OBJ(artifact:1)
CZ: {vyrábět, vyrobit} kdo*AG(person:1)| co*ACT(plant:1) =
co*OBJ(artifact:1)

uproot, eradicate, extirpate, exterminate – destroy completely, as if down to
the roots; "the vestiges of political democracy were soon uprooted"
BG: {izkorenjavam, premachvam} njakoj*AG(person:1)| neščo*AG(institution:2)
= neščo*ATTR(evil:3)|*EVEN(terrorism:1)
CZ: {vykořenit, vyhladit, zlikvidovat} kdo*AG(person:1)|co*AG(institution:2) =
co*ATTR(evil:3)|EVEN(terrorism:1)

carry, pack, take – have with oneself; have on one's person
BG: {nosja, vzimam} njakoj*AG(person:1)= neščo*OBJ(object:1)
CZ: {vzít si s sebou, brát si s sebou, mít s sebou, mít u sebe}
kdo*AG(person:1)= co*OBJ(object:1)

FIGURE 1: Common verb frame examples for Czech and Bulgarian

The construction of the valency frames for the Bulgarian verbs was performed in two stages:

1. Construction of the frames for those Bulgarian verb synsets that have corresponding (via Interlingual Index number) verb synsets in the CzWN and in addition these CzWN synsets are provided with already developed frames.
2. Creation of frames for verb synsets without analogues in the CzWN. The frames for more than 500 Bulgarian verb synsets have been created and the overall number of added frames was higher than 700. About 25% of the Bulgarian verb valency frames we used without any changes, they match the Czech ones completely.

In our view the Bulgarian experiment is convincing enough and it shows sufficiently that it is not necessary to create the valency frames for the individual languages separately.

4.2 Romanian example

(Tufis *et al.*, 2006) investigated the feasibility of the importing the valency frames defined in the Czech WordNet (Pala and Smrž, 2004) into the Romanian WordNet. They simply attached Czech valency frames from Czech WordNet to the Romanian verbs. As we hinted above the Czech CVFs specify syntactic and semantic restrictions on the predicate argument structure of the predicate denoting the meaning of a given synset. Let us consider, for instance, the Romanian verbal synset ENG20-02609765-v (a_se_afla:3.1, a_se_g'asi:9.1, a_fi:3.1) with the gloss "be located or situated somewhere; occupy a certain position." Its valency frame is described by the following expression:(nom*AG(fiint'a:1.1)— nom*PAT(obiect_fizic:1)) = prep-acc*LOC(loc:1).

The specified meaning of this synset is: an action the logical subject of which is either a fiint'a (sense 1.1) with the AGENT role(AG), or a obiect_fizic (sense 1) with the PATIENT role (PAT). The logical subject is realized as a noun/NP in the nominative case (nom). The second argument is a loc (sense 1) and it is realized by a prepositional phrase with the noun/NP in the accusative case (prep-acc). Via the interlingual equivalence relations among the Czech verbal synsets and Romanian synsets about 600 valency frames were imported. They were manually checked against the BalkaNet test-bed parallel corpus (Erjavec *et al.*, 2004) and more than 500 complex valency frames were found valid as they were imported or with minor modifications. This result supported by the real evidence confirms well our previous assumptions. Czech CVFs also motivated Tufis' group for further investigation on automatically acquiring FrameNet structures for Romanian and associating them with WordNet synsets. The fact that Romanian has only five cases in comparison with 7 in Czech did not meant a complication in the experiment.

Recently, a similar experiment has been started for building the Czech version of FrameNet (Materna, 2009). It appears that the subcategorization features used in CVFs can be linked with slots in FrameNet reasonably well. In doing this we also take advantage of the fact that Czech WordNet is linked to PWN v.2.0 through ILI.

4.3 English example

Let us take the complex valency frame for the Czech verb *učit se (learn)* and its deep valency part describing the basic meaning of the verb:

```
kdo1*AG(person:1)=co4*KNOW(information:3)[kde]*GROUP(institution:1)
```

(ex.: *učit se matematiku ve škole – to learn mathematics in the school*)

According to the principle of translatability the translation pair *učit se – learn* can be considered correct. Thus we can conclude that this particular frame can work well both for Czech and English. Similarly, take the Czech and English verb *pít/drink* with their basic meaning again. The relevant deep part of the CVF takes the following shape:

```
kdo1*AG((person:1)|(animal:1))=co4*SUBS(beverage:1)
```

(ex.: *bratr pije pivo, kůň pije vodu – my brother drinks beer, the horse drinks water*)

Again, it can be seen that this CVF describes well both Czech verb meaning and the meaning of its English translation equivalent.

It may be argued that more examples are needed and there may be some doubtful cases. However, at the moment we have about 8800 Czech verbs with their CVFs linked to the corresponding English verb synsets in PWN v.2.0 via ILI, thus relying on the principle of translatability we have enough examples in which CVFs can serve for Czech and English verbs equally well. Fortunately, there is also independent evidence that comes from the semantic classes of Czech and English verbs as they exist in Czech lexical database VerbaLex and partly VerbNet (Kipper *et al.*, 2000) (see below), for instance classes including verbs of drinking, eating, verbs denoting animal sounds, putting, weather and others (altogether 82 classes). Even brief comparison shows that their CVFs appear suitable for both languages and not only for them.

In VerbaLex we presently have about 10 500 Czech verb lemmata. From them only 5158 have been linked to the Princeton WordNet 2.0 via ILI in the first phase. After processing all VerbaLex verbs we have linked to Princeton WordNet further 3686 Czech verbs, i.e. 8844 Czech verbs are now linked to PWN v.2.0. The processing of the VerbaLex verbs and their linking to PWN v.2.0 shown, however, that approx. 15% of the Czech verb synsets cannot be linked to PWN v.2.0 straightforwardly since it is not possible to find their lexicalized translation equivalents in English. To be able to translate them to English the corresponding non-lexicalized English descriptions have to be found as the translations in the same way as translators usually do it.

5 Semantic Classes of Czech Verbs

We have worked out semantic classes of Czech verbs that were originally inspired by Levin's classes (Levin, 1993) and VerbNet classes (Kipper *et al.*, 2000). Since Czech is a highly inflectional language the patterns of alternations typical for English cannot be straightforwardly applied – Czech verbs require noun phrases in morphological cases (there are 7 of them both in singular and plural) and the category of aspect is grammatical in Czech. However, classes similar to Levin's can be constructed for Czech verbs as well but they have to be based on the grouping of the verb meanings. Before starting the VerbaLex project we had compiled a Czech-English dictionary with Levin's 49 semantic classes and their Czech equivalents containing approx. 3000 Czech verbs as the starting point.

In VerbaLex project we went further and associated CVFs of Czech verbs with the verb classes in a similar way as it was done in VerbNet. This meant that for each Czech verb in VerbaLex we marked the corresponding VerbNet semantic class a verb belongs to. Then we looked at the semantic roles occurring within the individual CVFs. This inevitably brought about the reduction of the VerbNet semantic classes from about 400 to 89 – the semantic roles helped us to make the semantic classification of the verbs more consistent. For example, take the label beverage:1 – it is yielding a homogeneous group containing 62 verbs. It can be seen that Levin's classes contain verbs that seem to form one consistent group but if we look at them closer it becomes obvious that they inevitably call for further differentiation and subclassification. For instance, if we take the class containing verbs of putting (PUT-9 in VerbaLex notation) we can see that it contains verbs like *to put* on one hand, but also *to cover* or *to hang* on the other. These differences in their meaning have to be captured by further subclassification in which the semantic roles play relevant role.

The basic assumption in this respect is that there is a mutual relation between the semantic classes of verbs and the semantic roles in their corresponding CVFs. In this way both the consistency of the inventory of semantic roles and consistency of the related semantic verb classes can be checked – obviously, in one class we can expect the roles specific only for that class. For example, for verbs of clothing the role like ART(garment:1) with possible subcategorizations reliably predicts the respective classes and their subclasses. Similarly it works for other verb classes, such as verbs of eating (role SUBS⟨food:1⟩), drinking (role SUBS⟨beverage:1⟩), emotional states (role FEEL⟨emotion:1⟩, weather (role PHEN⟨weather:1⟩ and others.

In our view, the good news also is that if the semantic part of the CVFs can work for more languages as we tried to show the same can be extended for the corresponding semantic verb classes.

The ultimate goal is to obtain semantic verb classes suitable for further computer processing and applications.

6 Conclusions

In the paper we have concentrated on the description of the background ideas behind the lexical database of Czech verbs VerbaLex whose main contribution consists in the development complex valency frames (CVFs) capturing the surface (morphological) and deep (semantic) valencies of the corresponding verbs. For labeling the roles in the valency frames we have worked out a list (in fact an ontology) of the two-level semantic labels which at the moment contains approx. 40 'ontological' roles and about 600 subcategorization features represented by the literals taken from Princeton WordNet 2.0. At present VerbaLex contains approx. 10 500 Czech verbs with 19 000 CVFs.

Further, we pay attention to some multilingual implications and show that originally 'Czech' Complex Valency Frames can reasonably describe semantics of the predicate argument structures of Bulgarian, Romanian and English verbs and obviously also verbs in other languages. What has to be dealt with separately are surface valencies because they heavily depend on the morphological cases in Czech, Romanian and to some extent Bulgarian and syntactic rules of Romanian and English. The issue calls for further testing and validation, however, we consider the presented analysis more than promising.

Acknowledgments

This work has been partly supported by the Ministry of Education of CR within the National Research Programme II project 2C06009 and project LC536.

References

H. C. Boas, E. Ponvert, M. Guajardo, and S. Rao (2006), The current status of German FrameNet, in *SALSA workshop at the University of the Saarland*, Saarbrucken, Germany.

T. Erjavec *et al.* (2004), MULTEXT-East: Multilingual Text Tools and Corpora for Central and Eastern European Languages, `http://nl.ijs.si/ME/`.

C.J. Fillmore, C.F. Baker, and H. Sato (2004), FrameNet as a 'net', in *Proceedings of Language Resources and Evaluation Conference (LREC 2004)*, volume 4, pp. 1091–1094, ELRA, Lisbon.

P. Hanks (2004), Corpus Pattern Analysis, in *Proceedings of the Eleventh EURALEX International Congress*, Universite de Bretagne-Sud, Lorient, France.

P. Hanks, K. Pala, and P. Rychlý (2007), Towards an empirically well-founded semantic ontology for NLP, in *Workshop on Generative Lexicon*, Paris, France.

D. Hlaváčková and A. Horák (2005), VerbaLex – New Comprehensive Lexicon of Verb Valencies for Czech, in *Proceedings of the Slovko Conference*, Bratislava, Slovakia.

K. Kipper, H. T. Dang, and M. Palmer (2000), Class Based Construction of a Verb Lexicon, in *AAAI-2000 17th National Conference on Artificial Intelligence*, Austin TX.

S. Koeva (2004), Bulgarian VerbNet, Technical Report part of Deliverable D 8.1, EU project Balkanet.

S. Koeva *et al.* (June 2004), Restructuring WordNets for the Balkan Languages, Design and Development of a Multilingual Balkan Wordnet Balkanet, Technical Report Deliverable 8.1, IST-2000-29388.

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

B. Lönneker (2004), Lexical databases as resources for linguistic creativity: Focus on metaphor, in *Linguistic Creativity Workshop*, Lisbon.

J. Materna (2009), Czech Verbs in Semantics of the FrameNet, in *Conference on Czech Formal Grammar*, Brno, in print.

K. Pala and P. Smrž (2004), Building the Czech Wordnet, in *Romanian Journal of Information Science and Technology*, volume 7, pp. 1–13.

M. Straňáková-Lopatková and Z. Žabokrtský (2002), Valency Dictionary of Czech Verbs: Complex Tectogrammatical Annotation, in *LREC 2002, Proceedings*, volume III, pp. 949–956, ELRA.

D. Tufis, V. B. Mititelu, L. Bozianu, and C. Mihaila (2006), Romanian WordNet: New Developments and Applications, in *Proceedings of the Third International WordNet Conference – GWC 2006*, pp. 336–344, Masaryk University, Brno, Jeju, South Korea.

P. Vossen, editor (1998), *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, Kluwer Academic Publishers, Dordrecht.