Can Complex Valency Frames be Universal?

Karel Pala and Aleš Horák

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

Abstract. This paper deals with the comprehensive lexicon of Czech verbs named VerbaLex. It 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 take the position that CVFs can be suitable for a description of the argument predicate structure not only of Czech verbs but also verbs in other languages, particularly Bulgarian, Romanian and English.

1 Introduction

Semantic role annotation is usually based on the appropriate inventories of labels for semantic roles (deep cases, arguments of verbs, functors, actants) describing argument predicate structure of verbs. It can be observed that the different inventories are exploited in different projects, e.g. Vallex [1], VerbNet [2], FrameNet [3], Salsa [4], CPA [5], VerbaLex [6].

With regard to the various inventories a question has to be asked: how adequately they describe semantics of the empirical lexical data as we can find them in corpora? From this point of view it can be seen that some of the inventories are rather 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 with them labels cognitively different entities – BEVERAGES with *drink* and FOOD with *eat*. If we consider verbs like *see* or *hear* we can observe similar differences not mentioning the fact that one can see anything. This situation can be improved if we use more detailed subcategorization features which, however, in other lexicons (e.g. in VerbNet or Vallex 1.0) are exploited only partially. If we are not able to discriminate the indicated semantic distinctions the use of the frames with such labels in realistic applications can hardly lead to the convincing and reliable results.

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.

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008, pp. 41–48, 2008. © Masaryk University, Brno 2008

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 it in VerbaLex. We have already hinted that "shallow" roles seem to be very similar to syntagmatic functions. At the same time it should be obvious that information that a person functions as an agent who performs an action is not only syntagmatic. 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 a two-level role consisting of the ontological part, i.e. SUBS(tance), 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[1] 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 are being created gradually along with building the necessary ontology. Thus for 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 [8].

2 VerbaLex and Complex Valency Frames

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

- morphological and syntactic features of constituents,
- two-level semantic roles,
- links to PWN and Czech WordNet hypero/hyponymic (H/H) hierarchy,
- differentiation of animate/inanimate constituents,
- default verb position,
- verb frames linked to verb senses,
- VerbNet classes of verbs.

[1] cf. [5] and also [7]

3 Role Annotation and EWN Top Ontology

Presently, our inventory contains about 30 general or ontological labels selected from the EuroWordNet Top Ontology (EWN TO), with some modifications, and the 2nd-level subcategorization labels taken mainly from the Set of Base Concepts introduced in EuroWordNet Project (1999). The 2nd-level labels (approx. 200) selected from the Set of Base Concepts (BCs) are more concrete and they 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 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. These sequences can be interpreted as detailed selectional restrictions.

Two-level labels contain ontological labels taken from EWN TO (about 30) that include roles like AGENT, PATIENT, INSTRUMENT, ADDRESSEE, SUBSTANCE, COMMUNICATION, ARTIFACT at the 1st level. The 2nd-level labels that are combined with them are literals from PWN 2.0 together with their sense number.

The nice property of the Czech valency frames is that the semantic restrictions are endogenous, i.e. they are specified in terms of other synsets of the same WordNet.

The notation allows us to handle basic metaphors as well. An example of CVFs for *drink/pít* may take the form:

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

4 Can CVFs be Universal?

The building VerbaLex database started during the EU project Balkanet (Balkanet Project, 2002) when about 1,500 Czech verb valency frames were included in Czech verb synsets. 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 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 [9,10]. 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.

The question then remains whether 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 well their semantics. VerbaLex is incorporated into Czech WordNet and through ILI also to PWN 2.0, thus we have the necessary translation pairs at hand. This also can be applied to other WordNets linked to PWN v.2.0. Thus we rely on the principle of translatability which means that the deep valencies developed for Czech verbs can be reasonably exploited also for English (see the Section 3). There is a problem with surface valencies which in English are based on the fixed order SVOMPT and on morphological cases in Czech but we consider this rather 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. 1,500 valency frames (cf. [11]). Since both languages (Czech and Bulgarian) are Slavonic we assumed 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).

The construction of the valency frames of 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

D. Tufis et al [10] investigated the feasibility of the importing the valency frames defined in the Czech WordNet [12] into the Romanian WordNet. They simply attached Czech valency frames from CzWn 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 produce, make, create – create or manufacture a man-made product

BG: {proizveždam} njakoj*AG(person:1)| nešĉo*ACT(plant:1)= nešĉo*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)| nešĉo*AG(institution:2)= nešĉo*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)= nešĉo*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)

Fig. 1. Common verb frame examples for Czech and Bulgarian

given synset. The valency frames also specify the morphological cases of the arguments. 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 we imported about 600 valency frames. They were manually checked against the BalkaNet test-bed parallel corpus (1984) 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 is more than promising. Czech CVFs also motivated Tufis’ group for further investigations on automatically acquiring FrameNet structures for Romanian and associating them with WordNet synsets.

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*)

If the translation pair *učit se – learn* is correct then we can conclude that this frame is suitable 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 CVFs 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 these are only two examples and there may be some doubtful cases. When linking Czech and English verb synsets via ILI is finished more examples can be adduced to show that the CVFs can serve for Czech and English equally well not speaking about other languages. To get more necessary evidence we are going to examine first selected semantic classes of the Czech and English verbs (see the next section), such as verbs of drinking, eating, verbs denoting animal sounds, putting, weather. 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 approx. only 5,158 have been linked to the Princeton WordNet 2.0 via ILI earlier. After processing all VerbaLex verbs we have linked to Princeton WordNet further 3,686 Czech verbs. Altogether 8,844 Czech verbs are now linked to Princeton WordNet. 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 since it is not possible to find their lexicalized translation equivalents in English. It should be remarked that this is a serious problem which, however, has to be solved separately.

5 Semantic Classes of Czech Verbs

We have worked out semantic classes of Czech verbs that were inspired by Levin's classes [13] and VerbNet classes [2]. 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). However, classes similar to Levin's can be constructed for Czech verbs as well but they have to be based only 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. 3,000 Czech verbs.

In VerbaLex project we went further and tried to link Czech verbs with the verb classes as they are used in VerbNet – they are also based on Levin's classification extended to almost 400 classes. This meant that for each Czech verb in VerbaLex we had marked the VerbNet semantic class a verb belongs to. The next step, however, was to have a look at the semantic roles introduced in VerbaLex. This led us to the reduction of the VerbNet semantic classes to 89 – the semantic roles helped us to make the semantic classification of the verbs more consistent. For example, if we take semantic role BEVERAGE – it is yielding a homogeneous group containing 62 verbs. It can be seen that Levin's classes sometimes 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 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.

The basic assumption in this respect is that there is a mutual relation between 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, with verbs of clothing the role like GARMENT with its respective subcategorizations reliably predicts the respective classes and their subclasses. Similarly it works for other verb classes, such as verbs of eating (FOOD), drinking (BEVERAGE), emotional states, weather and many others.

In our view, the good news also is that if the semantic parts 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 deal with 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 (ontology) of the two-level semantic labels which at the moment contains approx. 30 'ontological' roles and 200 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. From them

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 and 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 Academy of Sciences of Czech Republic under the project 1ET100300419 and by the Ministry of Education of CR in the National Research Programme II project 2C06009 and project LC536.

References

1. Straňáková-Lopatková, M., Žabokrtský, Z.: Valency Dictionary of Czech Verbs: Complex Tectogrammatical Annotation. In: LREC 2002, Proceedings. Volume III, ELRA (2002) 949–956.
2. Kipper, K., Dang, H.T., Palmer, M.: Class Based Construction of a Verb Lexicon. In: AAAI-2000 17th National Conference on Artificial Intelligence, Austin TX (2000).
3. Fillmore, C., Baker, C., Sato, H.: Framenet as a 'net'. In: Proceedings of Language Resources and Evaluation Conference (LREC 2004). Volume 4., Lisbon, ELRA (2004) 1091–1094.
4. Boas, H.C., Ponvert, E., Guajardo, M., Rao, S.: The current status of German FrameNet. In: SALSA workshop at the University of the Saarland, Saarbrucken, Germany (2006).
5. Hanks, P.: Corpus Pattern Analysis. In: Proceedings of the Eleventh EURALEX International Congress, Lorient, France, Universite de Bretagne-Sud (2004).
6. Hlaváčková, D., Horák, A.: VerbaLex – New Comprehensive Lexicon of Verb Valencies for Czech. In: Proceedings of the Slovko Conference, Bratislava, Slovakia (2005).
7. Hanks, P., Pala, K., Rychlý, P.: Towards an empirically well-founded semantic ontology for NLP. In: Workshop on Generative Lexicon, Paris, France (2007) in print.
8. Vossen, P., ed.: EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers, Dordrecht (1998).
9. Koeva, S.: Bulgarian VerbNet. Technical Report part of Deliverable D 8.1, EU project Balkanet (2004).
10. Tufis, D., Barbu, Mititelu, V., Bozianu, L., Mihaila, C.: Romanian WordNet: New Developments and Applications. In: Proceedings of the Third International WordNet Conference – GWC 2006, Jeju, South Korea, Masaryk University, Brno (2006) 336–344.
11. Koeva, S., et al.: Restructuring wordnets for the balkan languages, design and development of a multilingual balkan wordnet balkanet. Technical Report Deliverable 8.1, IST-2000-29388 (June 2004).
12. Pala, K., Smrž, P.: Building the Czech Wordnet. In: Romanian Journal of Information Science and Technology. 7(2–3), (2004) 79–88.
13. Levin, B., ed.: "English Verb Classes and Alternations: A Preliminary Investigation". The University of Chicago Press, Chicago (1993).