Building a Large Lexicon of Complex Valency Frames

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

Abstract

This paper describes the process of building and using a new comprehensive lexicon of Czech verb valency frames based on *complex valency frames*. The main features of the lexicon entries are designed to bring important semantic information to computer processing of predicate constructions in running texts. The most notable features include two-level semantic labels with linkage to the Princeton and EuroWordNet hierarchy and surface verb frame patterns used for automatic syntactic analysis. Some implications for other languages, particularly English, Bulgarian and Romanian, are reported.

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 (Stranakova-Lopatkova and Zabokrtsky, 2002), VerbNet (Kipper et al., 2000), FrameNet (Fillmore et al., 2006), 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 as we can find them in corpora? From this point of view it can be seen that some of the inventories are more syntactic than semantic (e.g. Vallex 1.0). 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. Then the role PATIENT has to be regarded as mainly syntagmatic though using subcategorization features can improve the situation, however, usually they are not exploited in other lexicons (e.g. in Vallex 1.0). If we are not able to discriminate the indicated semantic distinctions the use of the frames with such labels in realistic applications may not lead to convincing and reliable results.

These considerations led us to the design of 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. 11 000 Czech verbs.

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, see below. We 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 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,

Figure 1: An example of a Complex Valency Frames for the verbs klesnout:1, klesat:1, padnout:1, padat:1, snést se:1, snášet se:1 (descend:1, fall:2, go down:1, come down:1).

```
who_nom*AGENT(human:1|animal:1) <eat:1/jíst:1> what_acc*SUBS(food:1)
     withwhat_ins*INS(cutlery:2)
who_nom*AGENT(human:1|animal:1|institution:1) <see:1/vidět:1> what_acc*ANY(anything:1)
who_nom*AGENT(human:1|animal:1) <hear:1/slyšet:1> what_acc|koho4*PHEN(sound:1)
     how*MAN(manner:1).
```

Figure 2: Translation of Czech CVFs to English.

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 which are "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, that are partly based on the TOP Ontology as used in EuroWordNet project (Vossen, 1998). For semantic roles and types Brandeis Shallow Ontology ((Pustejovsky et al., 2006)) has been used but it is not regarded a final solution at the moment. (Examples of the semantic roles and types can be found in the papers quoted above.)

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 suitable for syntactic and semantic analysis. After reviewing actual verb frame repositories, we have developed *Complex Valency Frames* (CVFs) that contain:

- morphological and syntactic features of constituents
- two-level semantic roles
- links to PWN and Czech WordNet hyper/hyponymic (H/H) hierarchy
- differentiation of animate/inanimate constituents

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 3: Common verb frame examples for Czech and Bulgarian

- default verb position
- verb frames linked to verb senses
- VerbNet classes of verbs.

An example of a CVF is displayed in the Figure 1.

3 Role Annotation and EWN Top Ontology

Presently, our inventory contains the general or ontological labels selected from the EuroWordNet Top Ontology (EWN TO), with some modifications, and the 2^{nd}-level subcategorization labels taken mainly from the Set of Base Concepts introduced in (EuroWordNet Project, 1999). The 2^{nd}-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 coming from EWN TO. The motivation for this choice is based on the fact that WordNet has a hierarchical structure which covers about 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 quite detailed selectional restrictions.

The two-level labels contain ontological labels taken from EWN TO (about 40) that include roles like AGENT, PATIENT, INSTRUMENT, ADDRESSEE, SUBSTANCE, COMMUNICATION, ARTIFACT at the 1^{st} level. The 2^{nd}-level labels that are combined with them are literals from PWN 2.0 together with their sense number.

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

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

4 Multilingual Aspects of CVFs – can CVFs be Universal?

We have started building VerbaLex database during the EU project Balkanet (Balkanet Project, 2002) when about 1500 Czech verb valency frames were included in Czech WordNet. They were linked to English and other languages within Balkanet through the Interlingual Index (ILI). In the Balkanet project an experiment took place in which CVFs developed for Czech verbs have been linked to the corresponding verbs of Bulgarian and Romanian (Koeva, 2004).

While the experience with Czech CVFs for Bulgarian and Romanian is positive (see below the Section 4.1), and the result can be generalized also for other Slavonic languages like Slovak or Polish, the question remains whether CVFs developed for Czech can be applied to English as well. If we exploit ILI and have look at the VFs for Czech/English

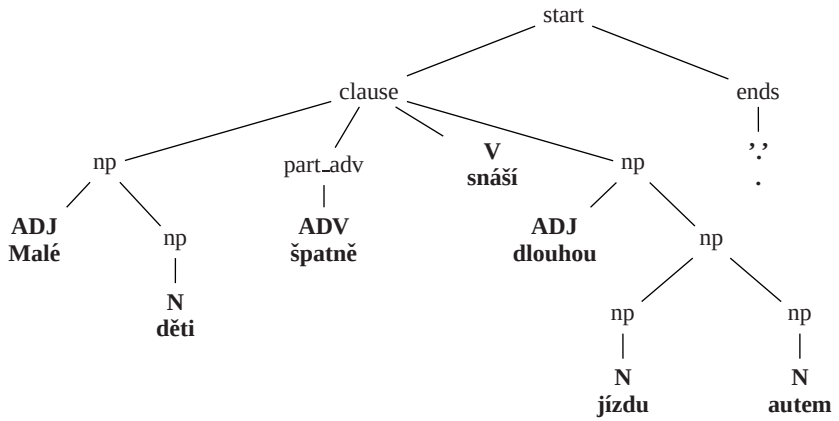

Figure 4: Syntactic tree of an example input sentence “`Malé děti špatně snáší dlouhou jízdu autem.`” (Small children badly withstand long journey by car.)

verbs like *pít/drink, jíst/eat* and apply them to their English translation equivalents we come to the conclusion that the Czech deep valencies certainly can describe their semantics. This conclusion is based on the simple assumption that we have the correct translation equivalents at our disposal. 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 for other WordNets linked to PWN. If the principle of translatability holds it means that the deep valencies developed for Czech can be reasonably exploited also for English (see the Figure 2).

In our view, the roles designed originally for the Czech verbs can serve for the corresponding English equivalents as well.

4.1 Bulgarian example

The enrichment of Bulgarian WordNet with verb valency frames was initiated by the experiments with Czech WordNet (CzWN) which already contained approx. 1500 valency frames (cf. (Koeva and others, June 2004)). Since both languages (Czech and Bulgarian) are Slavonic we assumed that a relatively great part of the verbs should realize their valency in the same way. The examples of Bulgarian and Czech valency frames in the Figure 3 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 completely coincide with the Czech ones.

Similar results have been obtained also for Romanian where a good agreement was observed on the semantic level but the surface valencies had to be re-processed, Czech and Romanian are morphologically different.

In our view these experiments are convincing enough and they show sufficiently that it is not necessary to create the valency frames for the individual languages separately.

4.2 Levin’s Classes and Czech Verbs

We have created semantic classes of Czech verbs that are inspired by Levin’s classes (Levin, 1993) and VerbNet classes (Kipper et al., 2000). Since Czech is a highly inflectional language the patterns of alternation typical for English cannot be straightforwardly applied – Czech verbs require noun phrases in the morphological cases (there are 7 of them both in singular and plural). However, classes similar to Levin’s can be constructed for

```
verb_rule_schema:  3 nterms, '#2'
   nterm 1:  k1gNnPc1
   nterm 2:  k5eAp3nPtPmIaI
   nterm 3:  k1gFnSc4
   group 1:  0, 2, +npnl -> .{ left_modif } np .  k1gMnSc1 ''malé děti''
   group 2:  2, 3, +ADV -> .'špatně' .  k6xMeAd1
   group 3:  4, 7, +npnl -> .{ left_modif } np .  k1gFnSc4 ''dlouhou jízdu autem''
possible subjects:  #1
Clause valency list:
     snášet <v>-#2:(1)hH#1:(0)hPTc1-#3:(2)hPTc4
     snášet(0) <v>#1:(1)hH-#2:(2)hPTc4
Verb valency list:
     snášet <v>#2:hH-#1:hPTc4
     snášet <v>#1:hPTc4
Matched valency list:
     snášet(0) <v>#2:(1)hH-#1:(2)hPTc4
```

Figure 5: The output of the verb frame extraction algorithm during the example sentence analysis.

Czech verbs as well but they have to be based only on the semantics of the verb classes. Before the starting the VerbaLex project we had compiled a Czech-English dictionary with Levin's 50 semantic classes and their Czech equivalents containing approx. 3000 Czech verbs.

In VerbaLex project we went further and linked Czech verbs with the verb classes as they are used in VerbNet – they are also based on Levin's classification extending it to almost 400 classes. This means that for each Czech verb in VerbaLex we mark the VerbNet semantic class a verb belongs to. We consider this information useful though it is known (according to our knowledge at least) that Levin's classes have not been extensively confronted with any corpus data. This certainly makes them less reliable.

The basic assumption in this respect is that the semantic classes of verbs should be helpful in checking the consistency of the inventory of semantic roles since in one class we can expect the roles specific only for that class. For example, with verbs of clothing the role like GARMENT and its respective subcategorizations can be reliably predicted, similarly it should work for other verb classes, such as verbs of eating, drinking, wearing, emotional states, weather and others. In the close future we plan to compare VerbNet semantic classes with the classes that we expect to obtain by sorting our valency frames according to the roles they occur with.

5 Application in Syntactic Analysis

We are currently testing the application in our syntactic analyzer `synt` that is designed for parsing real-text sentences. The verb frame extraction (VFE) process in `synt` is controlled by the metagrammar semantic actions. The parser builds a forest of values[1] to represent a result of the application of contextual constraints. The VFE actions are then executed on a different level (Horák and Kadlec, 2005) than the "usual" actions, which allows us to apply VFE actions on the whole forest of values.

If the analyzed verb has a corresponding entry in VerbaLex, we try to match the extracted frame with frames in the lexicon. When checking the valencies with VerbaLex, the dependence on the surface order is discharged. Before the system confronts the actual verb valencies from the input sentence with the list of valency frames found in the lexicon, all the valency expressions are reordered. By using the standard ordering of participants, the valency frames can be handled as sets independent on the current position of verb arguments. However, since VerbaLex contains an information about the *usual* verb position within the frame, we promote the standard ordering with increasing or decreasing the respective derivation tree probability.

The system processing can be presented on an example sentence – see the syntactic tree in the Figure 4 and the textual output of the part of the system

[1] a DAG (directed acyclic graph) structure that corresponds to the resulting chart structure supplemented with values computed during the semantic actions like feature agreement tests or verb frame extraction

that works on the VFE algorithm in the Figure 5. The system first identifies the verb rule constituents (`nterm`s), then the corresponding groups, i.e. the actual sentence constituents that will play the role as verb frame arguments, are extracted from the forest of values. Groups usually do not correspond to nterms one-to-one, since they are stored within nonterminals deeper in the forest and not directly in the verb rule. This part of the VFE algorithm has unfortunately exponential time complexity, however, for common sentences the depth of the verb frame constituents is not more than three levels, so the actual running times are usually within fractions of seconds. After the identification of the groups, the algorithm looks for possible subjects – this is not as easy as it may look at the first sight, since the sentence subject can be expressed not only by a noun phrase in nominative (which is the most frequent option in Czech), but also by e.g. prepositional phrase or verb infinitive. If no possible subject is found, the algorithm supplies a pronoun for an inexplicit subject with the gender corresponding to the verb. The Clause valency list displays all possible combinations of the translations of the verb arguments found into verb frame patterns. This list is then intersected with the list of lexicon entries for the verb to obtain the Matched valency list as a result of the VFE algorithm.

The effectiveness of the syntactic analysis with the VFE algorithm was measured on approximately 4.000 Czech corpus sentences with the median of 15 words per sentence and the Clause valency list contained 11 possible verb frames with the running time of 0.07 seconds per sentence.

6 Conclusions

In the paper we report on the building the lexical database of Czech verbs VerbaLex with their surface (morphological) and deep (semantic) valencies. For labeling the roles in the valency frames we have developed a list (ontology) of the two-level labels which at the moment contains approx. 40 'ontological' roles and 200 subcategorization features represented by the literals taken from Princeton WordNet 2.0. At present VerbaLex contains approx. 11 000 Czech verbs with 28 000 frames. We also mention some multilingual implications and show how the CVFs can be exploited in syntactic analysis of Czech.

Acknowledgments

This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536 and in the National Research Programme II project 2C06009 and by the Czech Science Foundation under the project 201/05/2781.

References

Hans C. Boas, Elias Ponvert, Mario Guajardo, and Sumeet Rao. 2006. The current status of German FrameNet. In *SALSA workshop at the University of the Saarland*, Saarbrucken, Germany.

C.J. Fillmore, C.F. Baker, and H. Sato. 2006. Framenet as a 'net'. In *Proceedings of Language Resources and Evaluation Conference (LREC 04)*, volume vol. 4, 1091-1094, Lisbon. ELRA.

Patrick Hanks, Karel Pala, and Pavel Rychlý. 2007. Towards an empirically well-founded semantic ontology for NLP. In *Workshop on Generative Lexicon*, Paris, France. in print.

Patrick Hanks. 2004. Corpus pattern analysis. In *Proceedings of the Eleventh EURALEX International Congress*, Lorient, France. Universite de Bretagne-Sud.

Dana Hlaváčková and Aleš Horák. 2005. Verbalex – new comprehensive lexicon of verb valencies for czech. In *Proceedings of the Slovko Conference*, Bratislava, Slovakia.

Aleš Horák and Vladimír Kadlec. 2005. New meta-grammar constructs in Czech language parser synt. In *Proceedings of Text, Speech and Dialogue 2005*, pages 85–92, Karlovy Vary, Czech Republic. Springer-Verlag.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class based construction of a verb lexicon. In *AAAI-2000 Seventeenth National Conference on Artificial Intelligence*, Austin TX.

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.

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

Beth Levin, editor. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago Press.

J. Pustejovsky, C. Havasi, R. Sauri, and P. Hanks. 2006. Towards a generative lexical resource: The brandeis semantic ontology. In *Language Resources and Evaluation Conference, LREC 2006*.

M. Stranakova-Lopatkova and Z. Zabokrtsky. 2002. Valency dictionary of czech verbs: Complex tectogrammatical annotation. In *LREC2002, Proceedings*, volume III, pages 949–956. ELRA.

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