VerbaLex – New Comprehensive Lexicon of Verb Valencies for Czech

Dana Hlaváčková and Aleš Horák

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

Abstract. The paper presents new lexicon of verb valencies for the Czech language named VerbaLex. VerbaLex is based on three valuable language resources for Czech, three independent electronic dictionaries of verb valency frames.
The first resource, Czech WordNet valency frames dictionary, was created during the Balkanet project and contains semantic roles and links to the Czech WordNet semantic network. The other resource, VALLEX 1.0, is a lexicon based on the formalism of the Functional Generative Description (FGD) and was developed during the Prague Dependency Treebank (PDT) project. The third source of information for VerbaLex is the syntactic lexicon of verb valencies denoted as BRIEF, which originated at FI MU Brno in 1996.
The resulting lexicon, VerbaLex, comprehends all the information found in these resources plus additional relevant information such as verb aspect, verb synonymity, types of use and semantic verb classes based on the VerbNet project.

1 Introduction

The beginnings of building the verb valency frame dictionary at the Faculty of Informatics at Masaryk University (FI MU) dates back to 1997 [1]. Since then, the dictionary, denoted as Brief, has undergone a long development and has been used in various tools from semantic classification to syntactic analysis of Czech sentence [2]. Currently, the dictionary plays a key role within an experimental high-coverage syntactic analysis using the data from the Czech WordNet. The data in this dictionary can be entered in several mutually convertible formats:

brief:

```
jíst (to eat) <v>hTc4,hTc4-hTc6r{na}, hTc4-hTc7
```

verbose:	display:
jíst	
= co	jíst něco
= co & na čem	jíst něco na něčem
= co & čím	jíst něco něčím

Lemma variants:

Princeton WordNet – plan:2
Definition: make plans for something
VALLEX 1.0: vymyslet$_2$ / vymyslit$_2$
VerbaLex: vymyslet:1, vymyslit:1, naplánovat:3

Word entries:

Princeton WordNet – arrive:1, get:5, come:2
Definition: reach a destination; arrive by movement or progress
VALLEX 1.0: dojít$_1$
VerbaLex: dojít:1, dorazit:1, dostat se:1, přicestovat:1, přijet:1, přijít:1

Fig. 1. Examples of verb frame entry heads for verbs with lemma variants and for synonymic verbs.

The Brief dictionary contains about 15 000 verbs with 50 000 verb valency frames, thus making it an invaluable language resource with high coverage. However, the different verb senses are not distinguished here.

Another advance in the Czech verb valency processing came during the work on the Czech WordNet within the Balkanet project [3]. The Czech WordNet has been supplemented with a new language resource, Czech WordNet valency frames dictionary. The new acquisition of this dictionary were the semantic roles and links to the Czech WordNet semantic network.

During the work on enhancing the list and adding new entries into it, we have come to the need of comparing the quality and features of the list with the parallelly created valency lexicon of Czech verbs denoted as VALLEX 1.0 [4]. In cooperation with the VALLEX team, valency frames from Czech WordNet were transformed to an augmented VALLEX format, which was named VerbaLex.

The FI MU VerbaLex dictionary is being actively developed, checked and supplemented with new data. Currently, VerbaLex contains 3 469 verb literals which, when gathered in synonymic groups, share 1 807 verb frames. Nowadays, several linguists are working on a bulk of 15 000 more verbs being added to VerbaLex.

2 Linguistic requirements for the VerbaLex format

In this section, we present the substantiation of the main differences between VerbaLex and VALLEX 1.0 valency frames notation.

The lexical units in WordNet are organized into synsets (sets of synonyms) arranged in the hierarchy of word meanings (hyper-hyponymic relations). VerbaLex differs from VALLEX 1.0 in augmentation of the original format, detailed differentiation of valency frames and above all semantic roles (deep cases). For that reason, the headwords in VerbaLex are formed with lemmata in a synonymic relation (synset subsets) followed by their sense numbers (standard Princeton

WordNet notation). The standard definition of synonymity says that two synonymic words can be always substituted in the context. However, the synonymity in synsets is understood like very close sense affinity of given words, the substitution rule cannot be applied in all cases here. In VALLEX 1.0, a headword is one lemma, possibly two or more lemmata in case of lemma variants.[1] Lemma variants in VerbaLex are considered as independent lemmata and they are distinguished by their WordNet sense numbers. An example of two verb frame entries in VALLEX 1.0 and VerbaLex is displayed in the Figure 1.

In VerbaLex, each word entry includes an information about the verb aspect (perfective – `pf.`, imperfective – `impf.` or both aspects – `biasp.`). VerbaLex valency frames are enriched with aspect differentiations for examples containing the verb used with the given valency frame. This is important in case of synonymic lemmata with different aspect:

Princeton WordNet – wade:1
Definition: walk (through relatively shallow water)
VerbaLex: brodit se:2 impf., přebrodit se:1 pf.
frame: AG <person:1>$^{\text{obl}}_{\text{kdo1}}$ VERB SUBS <substance:1>$^{\text{obl}}_{\text{čím7}}$
example: přebrodil se blátem pf. / he wade through mud
example: brodil se pískem impf./ he wade through sand

The constituent elements of frame entries are enriched with pronominal terms (e.g. `kdo` – who, `co` – what) and the morphological case number. This notation allows to differentiate an animate or inanimate agent position:

Princeton WordNet – bump:1, knock:3
Definition: knock against with force or violence
VerbaLex: narazit:1 pf. / narážet:1 impf.
frame: AG <person:1>$^{\text{obl}}_{\text{kdo1}}$ VERB OBJ <object:1>$^{\text{obl}}_{\text{do čeho2,na co4}}$
PART <body part:1>$^{\text{obl}}_{\text{čím7}}$
example: I bumped to the wall with my head
frame: OBJ <vehicle:1>$^{\text{obl}}_{\text{co1}}$ VERB OBJ <object:1>$^{\text{obl}}_{\text{do čeho2,na co4}}$
example: the car bumped to the tree

2.1 Verb usage and verb classes

VerbaLex captures additional information about types of verb use and semantic verb classes. Three types of verb use are displayed in the lexicon. The primary usage of a verb is marked with abbreviation *prim*, metaphorical use with *posun* and idiomatic and phraseological use with *idiom* (this follows the VALLEX 1.0 notation). The assigned semantic verb classes are adopted from the Martha Palmer's [5] VerbNet project. The verb classes list is based on Beth Levin's [6] classes with more fine-grained sets of verbs.

[1] the lemmata with small phoneme alternation that are interchangeable in any context without any change of the meaning – `bydlet/bydlit`, to live (where).

There are 395 classes in the current development version of VerbNet, which was provided by Martha Palmer's team. But this number seems to be too much for Czech verbs, therefore the list of verb classes will be adapted to the conditions of the Czech language:

Princeton WordNet – cry:2, weep:1
Definition: shed tears because of sadness, rage, or pain
VerbaLex: brečet:1, plakat:1, ronit:1
class: nonverbal_expression-40.2

Princeton WordNet – take care:2, mind:3
Definition: be in charge of or deal with
VerbaLex: dbát:2, starat se:2, pečovat:3
class: care-86

Princeton WordNet – be:11, live:5
Definition: have life, be alive
VerbaLex: žít:1, být:2, existovat:3
class: exist-47

3 Semantic roles

VerbaLex has introduced a different concept of semantic roles (i.e. functors in VALLEX 1.0) as compared to VALLEX 1.0. Currently, the list of semantic roles and the way of their notation establish one of the main differences between VALLEX 1.0 and VerbaLex valency frames (see also [7]). The functors used in VALLEX 1.0 valency frames seem to be too general and they do not allow distinguishing different senses of verbs. We suppose that a more specific subcategorization of the semantic role tags is necessary, therefore an inventory of two level semantic role labels was created.

The first level contains the main semantic roles proposed on the 1stOrderEntity and 2ndOrderEntity basis from EuroWordNet Top Ontology [8]. On the second level, we use specific literals (lexical units) from the set of Princeton WordNet Base Concepts with relevant sense numbers. We can thus specify groups of words (hyponyms of these literals) replenishable to valency frames. This concept allows us to specify valency frames notation with large degree of sense differentiability.

For example the literal `writing implement:1` is a hypernym for any implement that is used to write.

Princeton WordNet – draw:6
Definition: represent by making a drawing of, as with a pencil, chalk, etc. on a surface
VerbaLex: kreslit:1, malovat:1
frame: AG <person:1>$^{\text{obl}}_{\text{kdo1}}$ VERB ART<creation:2>$^{\text{obl}}_{\text{co4}}$ INS<writing implement:1>$^{\text{obl}}_{\text{čím7}}$
example: my sister draws a picture with coloured pencils, the famous artist was drawing his painting only with charcoal

The left-side valency position is most frequently occupied by the semantic role `AG`, an agent. The agent position in a valency frame is understood as a very general semantic role (functor `ACT`) in VALLEX 1.0. This label does not allow to distinguish various types of action cause. Two level semantic role labels in VerbaLex are able to define cause of action quite precisely. The main semantic role AG is completed by an adequate literal depending on the verb sense and valency frame. Thus, we can identify whether the agent is a person `AG(person:1)`, an animal `AG(animal:1)`, a group of people `AG(group:1)`, an institution `AG(institution:1)` or a machine `AG(machine:1)`. For some verbs with very specific sense, hyponyms of these literals are used. For example:

Princeton WordNet – sugar:1, saccharify:1
Definition: sweeten with sugar
VerbaLex: sladit:4, osladit:1, pocukrovat:1
frame: AG <person:1>$^{\text{obl}}_{\text{kdo1}}$ VERB SUBS <food:1>$^{\text{obl}}_{\text{co4}}$
SUBS <sugar:1>$^{\text{obl}}_{\text{čím7}}$
example: sugar your tea with brown sugar

In VALLEX 1.0, each valency frame starts always with functor ACT. In our opinion, it is useful to differentiate the sense of the left-side valency position (subject position) in more detail. According to our definition of agent AG (sb or sth doing sth actively) this position may be also occupied by other semantic roles. The subject position can contain objects OBJ, substances SUBS or a semantic role denoting abstract concepts – human activity ACT, knowledge KNOW, event EVEN, information INFO, state STATE. For example:

Princeton WordNet – follow:6, come after:1
Definition: come after in time, as a result
VerbaLex: přijít:25 / přicházet:25, následovat:4
frame: EVEN <event:1>$^{\text{obl}}_{\text{co1}}$ VERB EVEN <event:1>$^{\text{obl}}_{\text{po čem6}}$
example: heavy rain followed flood

Princeton WordNet – fall:3
Definition: pass suddenly and passively into a state of body or mind
VerbaLex: zachvátit:2, zmocnit se:2
frame: STATE <state:4>$^{\text{obl}}_{\text{co1}}$ VERB PAT <person:1>$^{\text{obl}}_{\text{koho4}}$
example: he fall into a depression

Quite a large number of semantic roles inspired by EuroWordNet Top Ontology roughly correspond with the PAT functor in VALLEX 1.0. The PAT label covers quite different senses, which can be very well identified.

In our inventory, PAT is defined as: the semantic role of an entity that is not the agent but is directly involved in or affected by the happening denoted by the verb in the clause (definition of literal `patient:2` from Princeton WordNet).

Princeton WordNet – experience:1, undergo:2, see:21, go through:1
Definition: go or live through

Table 1. List of semantic roles from VerbaLex that are used in examples.

AG	the semantic role of the animate entity that instigates or causes the happening denoted by the verb in the clause, we extended this definition for inanimate entity that does sth actively (e.g. machine)
ART	a man-made object taken as a whole
SUBS	that which has mass and occupies space
PART	a portion of a natural object, something determined in relation to something that includes it, something less than the whole of a human artifact
INS	a device that requires skill for proper use
OBJ	a tangible and visible entity; an entity that can cast a shadow
EVEN	something that happens at a given place and time
STATE	the way something is with respect to its main attributes

VALLEX 1.0: absolvovat$_2$
frame: $\text{ACT}_1^{\text{obl}}$ $\text{PAT}_4^{\text{obl}}$
VerbaLex: absolvovat:2, prožít:1 / prožívat:1 /
frame: AG <person:1>$_{\text{kdo1}}^{\text{obl}}$ VERB EVEN <experience:3>$_{\text{co4}}^{\text{obl}}$
example: he underwent difficult surgery

Some second level literals cannot be adopted from Princeton WordNet Base Concepts – especially specification of roles considered as "classic" deep cases. These literals (e.g. `agent:6`, `patient:2`, `donor:1`, `addressee:1` or `beneficiary:1`) do not have any hyponyms in Princeton WordNet and cannot be substituted by any word.

For such cases, the literal `person:1` is used (or another suitable literal with large number of hyponyms, e.g. `AG(person:1)`, `PAT(animal:1)`). This "classic" semantic roles are consistent with some functors in VALLEX 1.0 (`ACT`, `PAT`, `ADDR`, `BEN` etc.). A list of VerbaLex semantic roles that are used in the presented examples is displayed in the Table 1.

3.1 Special semantic roles

VerbaLex describes not only the valency and semantic frames, it also includes other relevant information about Czech verbs, such as the verb position. In a free-word order language like Czech the position of the verb within the verb frame is usually not strictly specified.

VerbaLex uses a special semantic role, `VERB`, which marks the (usual) position of the verb in its verb frame. Such default verb position is not needed only for analysis of verb valencies, it can be also directly used in the process of generation of Czech sentences, e.g. as an output of a question-answering machine.

The left side of the verb position is traditionally occupied by the sentence subject, which is also the case marked in most of the verb frames in VerbaLex. However, there are some cases, where the verb frame has to obey different rules

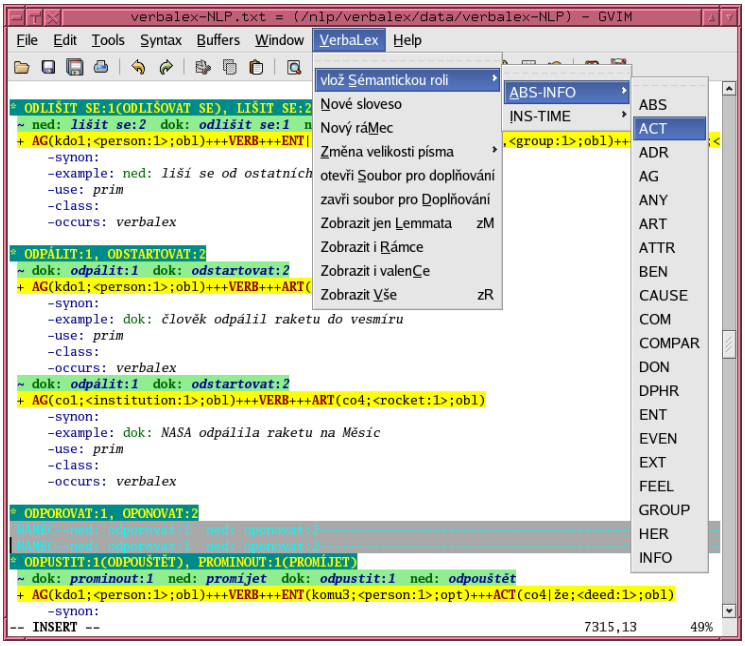

Fig. 2. The tool for editing verb valency frames dictionary in the VerbaLex format.

– e.g. sentence *Dalo se do deště* (It started to rain) cannot contain any subject. For the notation of such cases, VerbaLex uses another special semantic role `ISUB`, an inexplicit subject.

4 Implementation of editing and exporting tools

For the sake of editing the newly adopted verb valency frame format VerbaLex, we have implemented a new set of editing and exporting tools.

The main interactive tool for user editing of the valency dictionary, named `verbalex.sh`, is based on a highly configurable multi-platform editor VIM (see the Figure 2). Such approach enables a linguistic expert to easily enter computer-parseable data in a fixed plain text format and still, thanks to the flexible color syntax highlighting, he or she has a full visual control of possible errors in the format.

The editing itself is not fixed to one platform, users can run the same environment under any of the current popular computer operating systems (VIM editor runs on nearly any platform).

The authoring tool `verbalex.sh` currently offers these functions to the editing user:

– free editing of the dictionary entries

```
<headword_lemmata>
    <lemma ord='1' sense='2' aspect='pf'>chopit</lemma>
    <lemma ord='2' sense='2' aspect='pf'
        aspectual_counterpart_lemma='uchopovat'>uchopit</lemma>
    <lemma ord='3' sense='2' aspect='impf'
        aspectual_counterpart_lemma='uchopit'>uchopovat</lemma>
    <lemma ord='4' sense='3' aspect='pf'
        aspectual_counterpart_lemma='brát'>vzít</lemma>
    <lemma ord='5' sense='3' aspect='impf'
        aspectual_counterpart_lemma='vzít'>brát</lemma>
    <lemma ord='6' sense='4' aspect='pf'
        aspectual_counterpart_lemma='chápat se'>chopit se</lemma>
    <lemma ord='7' sense='4' aspect='impf'
        aspectual_counterpart_lemma='chopit se'>chápat se</lemma>
</headword_lemmata>
```

Fig. 3. An example of XML structure of aspectual counterpart tuples within one dictionary entry.

- regular expression searching in the dictionary
- template-based adding of a new verb entry or a new verb frame to the current entry
- menu-based adding of new semantic role to the current frame
- multilevel folding – hiding/unhiding of valency attributes, valencies or full valency frames
- visual marking of the current frame for further inquiry
- interactive merging of definitions from two parallel sources

Moreover, the interpreted approach of the tool makes adding of new features to the editing system easy to implement.

The plain text format edited by a human expert is in further processing transformed into an XML standard format which enables conversions into different formats used for visual checking, searching and presentation of the valency dictionary.

The XML schema used in VALLEX 1.0 had to be changed to suit the augmentation of the format in VerbaLex. The changes include

- adding `class` attribute to frame `slot` tag to cover wordnet basic concept literals
- including the wordnet word sense in the lemma tags
- shifting the verb aspect to `headword_lemma`, which now enumerates all the aspectual counterpart tuples. An example of such XML substructure can be found in the Figure 3.

The resulting XML structure is then transformed into various output formats with the use of modified tools from VALLEX 1.0. The export formats are HTML with navigation among the characteristic features of the dictionary

entries, Postscript document for printing including page index of all verbs and PDF, which allows navigation through the document in the same visual form as for hardcopy printing.

5 Conclusions and Future Directions

We have displayed the details of the VerbaLex verb valency frames dictionary and described the augmentation of the PDTB VALLEX 1.0 format that was needed for encapsulation of new semantic roles and links to the Czech wordnet entries.

The nearest development of VerbaLex dictionary includes adding several thousands of verbs and implementation of sophisticated checks of the correctness of the entered data with direct linking of the editing tool to wordnet editor and to the syntactic analyzer.

6 Acknowledgments

This work has been partly supported by Czech Science Foundation under the project 201/05/2781 and by Grant Agency of the Academy of Sciences of CR under the project 1ET400300414.

References

1. Karel Pala and Pavel Sevecek. Valence českých sloves (Valencies of Czech Verbs). In *Proceedings of Works of Philosophical Faculty at the University of Brno*, pages 41–54, Brno, 1997. Masaryk University.
2. P. Smrz and A. Horak. Determining type of TIL construction with verb valency analyser. In *Proceedings of SOFSEM'98*, pages 429–436, Berlin, 1998. Springer-Verlag.
3. Balkanet project website, `http://www.ceid.upatras.gr/Balkanet/`.
4. M. Stranakova-Lopatkova and Z. Zabokrtsky. Valency dictionary of czech verbs: Complex tectogrammatical annotation. In C. Paz Suárez Araujo M. González Rodríguez, editor, *LREC2002, Proceedings*, volume III, pages 949–956. ELRA, 2002.
5. Martha Palmer Joseph Rosenzweig Hoa Trang Dang, Karin Kipper. Investigating regular sense extensions based on intersective levin classes. In *Proceedings of Coling-ACL98*, Montreal CA, August 11-17, 1998. `www.cis.upenn.edu/~mpalmer/`.
6. Beth Levin, editor. *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago Press.
7. Karel Pala. Valency frames and semantic roles (in Czech). In *Proceedings of Slovko 2005 Conference*, Bratislava, 2005.
8. P. Vossen, L. Bloksma, et al. The EuroWordNet base concepts and top ontology. Technical Report Deliverable D017, D034, D036, WP5 EuroWordNet, LE2-4003, University of Amsterdam, 1998.