The Relations between Semantic Roles and Semantic Classes in VerbaLex

Dana Hlaváčková

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

Abstract. In this paper we present the database of verb valency frames for Czech language named VerbaLex being created presently in NLP Laboratory at Faculty of Informatics Masaryk University. This work involves building the valency database of Czech verbs with their surface and deep valency frames. Moreover we adopt the list of verb semantic classes from English to Czech. We want to show the way of more precisely subclassification of semantic classes for Czech verbs.

1 Introduction

VerbaLex – a large lexical database of Czech verb valency frames has been developed since 2005 at the Natural Language Processing Laboratory at the Faculty of Informatics Masaryk University (FI MU). VerbaLex is based on three existing independent resources:

1. BRIEF – dictionary of 50 000 valency frames for 15 000 Czech verbs is source of lexical data for VerbaLex. BRIEF was created at FI MU in 1997 [1]. The different verb senses are not distinguished here and valency frames are surface only, without any semantic information.
2. VALLEX – valency lexicon of Czech verbs is based on the formalism of the Functional Generative Description (FGD) and has been developed during the Prague Dependency Treebank (PDT) project [2]. Vallex and VerbaLex are similar projects with some important distinctions. The way of transformation of plain text format (for dictionary editing) to another formats (xml, pdf, html) used in Vallex has been used and changed for VerbaLex.
3. Czech WordNet valency frames dictionary, was created during the Balkanet project [3] and contains 1 359 valency frames (incl. semantic roles) associated with 824 sets of synonyms (synsets).

The organization of lexical data in VerbaLex comes out from the WordNet structure [4]. The lexical units in WordNet are organized into synsets arranged in the hierarchy of word meanings (hyper-hyponymic relations). 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 *basic valency frames* (BVF) display two types of information – the constituent elements of valency frames cover both syntactic level

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2007, pp. 97–101, 2007.

and lexical semantic level. The default verb position 'VERB' as the centre of the sentence is marked on the syntactic level. The pattern of sentence constituents are situated in left and right positions in accordance with the complementarity needed by the verb. The constituent elements of frame entries are entered as pure pronominal terms, e.g. kdo (who), co (what), or prepositional phrase pattern (with the lemma of the preposition) followed by the number of the required grammatical case of the phrase. This way of notation allows to differentiate an animate or inanimate subject or object position. The types of verbal complementation (nouns, adjectives, adverbs, infinitive constructions or subordinate clauses) are precisely distinguished in the verb frame notation. There is marked up the type of valency relation for each constituent element – obligatory 'obl' (must be present) or optional 'opt'. BVF is followed by simple example of usage verb in sentence. For example:
Synset: bavit:1, rozptýlit:2, rozptylovat:2
(PrincetonWordNet: amuse:2 /make (somebody) laugh)

frame: AG <person:1>$^{\text{obl}}_{\text{whoNom}}$ VERB PAT <person:1>$^{\text{obl}}_{\text{whatAccus}}$ ACT <act:2>$^{\text{opt}}_{\text{by doing whatInstr}}$

– *example: impf: bavil děti hrou (he amused the children by playing the game)*

VerbaLex captures additional information about the verbs which is organized in *complex valency frames* (CVF):

– definition of verb meanings for each synset;
– verb ability to create passive form;
– number of meaning for homonymous verbs;
– semantic classes;
– aspect (*perfective – pf., imperfective – impf. or both aspects – biasp.*);
– types of verb use (*primary – prim., figurative – fig., idiomatic – idiom.*);
– types of reflexivity for reflexive verbs.

For example:
SYNSET: BAVIT:1, ROZPTÝLIT:2, ROZPTYLOVAT:2
DEFINITION: poskytovat někomu zábavu/make (somebody) laugh

– *passive: yes*
– *meaning: I*
– *class: amuse-31.1-1*
– *impf: bavit:1 pf: rozptýlit:2 impf: rozptylovat:2*

frame: AG <person:1>$^{\text{obl}}_{\text{whoNom}}$ VERB PAT <person:1>$^{\text{obl}}_{\text{whatAccus}}$ ACT <act:2>$^{\text{opt}}_{\text{by doing whatInstr}}$

– *example: impf: bavil děti hrou (he amused the children by playing the game)*

– *attr: use: prim, reflexivity: obj_ak*

Current version of VerbaLex 2.0 contains 7 063 synsets, 23 461 verb senses, 10 596 verb lemmata and 21 100 valency frames. Valency database is available in txt, xml and html formats [5].

2 Semantic Roles

Semantic information of verb complementation is represented by two-level semantic roles in BVF. The first level contains the main semantic roles proposed on the 1stOrder-Entity and 2ndOrderEntity basis from EuroWordNet Top Ontology [6]. The 1^{st} level semantic roles represent close list of 29 semantic tags (e.g. *AG – agent, OBJ – object, INS – instrument, ACT – activity, INFO – information, SUBS – substance* etc.). On the second level, we use specific literals (lexical units) from the set of PrincetonWordNet 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 (e.g. *SUBS(beverage:1), OBJ(furniture:1), INS(edge tool:1)* etc.). The list of 2^{nd} level semantic roles is open, current version contains about 1 000 wordnet lexical unites.

3 Semantic Classes

We work with verb semantic classes that were originally adopted from the Levin's list of English verb classes [7] (48 classes). We also use the list of Martha Palmer's VerbNet project with more fine-grained sets of verbs [8] (82 classes, total of 395 subclasses). These verb classes have been translated and adopted for Czech language. Czech classes were enriched with Czech synonyms, aspect counterparts and Czech prefixed verbs. Presently, we work with 82 semantic verb classes, 258 subclasses and 6 393 Czech verb lemmata in the current version of our list. In building the semantic classes we prefer semantic criteria against the syntactic alternations used by Levin. As a result we get verb classes that are semantically more consistent than Levin's.

4 Relations

The process of adopting and enriching Czech semantic classes initially started with Levin/Palmer's classes but within VerbaLex we try to modify them with regard to the semantic features of predicate-argument structures of Czech verbs. Our aim is to create classes based also on the inventory of the semantic roles denoting verb arguments. This approach allows us to build semantic classes and subclasses more precisely in many cases.

Our point of view is based on assumption that verbs complemented by the identical 2^{nd} level semantic roles belong to one semantic class. For example, the

verbs linked to the semantic role beverage:1 (it occurs in 42 valency frames in VerbaLex) can create following semantic groups:

beverage consumption – *pít/drink, upíjet/sip, bumbat/guggle, ochutnávat/taste...*
oversized beverage consumption – *chlastat/booze, opíjet se/soak, přihnout si/swig...*
beverage serving – *čepovat/tap, točit/draw, nalévat/pour, napojit/water...*
beverage preparation – *zkvasit/ferment, vařit/brew, ledovat/frost, protřepat/shake...*
physical result after oversized beverage consumption – *zvracet/vomit, dávit/throw up, blinkat/be sick...*

In Levin/Palmer's list of semantic classes this type of verbs belongs mostly to class 39. Verbs of Ingesting and to wide and more closely undefined class 45. Verbs of Change of State.

Verbs complemented by 2^{nd} level semantic role *furniture:1* (it occurs in 60 valency frames in VerbaLex) can create following semantic groups:

furniture usage – *posadit se/sit down, ležet/lie, uložit se/lie down...*
furniture handling – *sklopit/recline, uklidit/tidy away, srovnat/order, umístit/place, stěhovat/move, otevřít/open...*
furniture making and maintenance – *čalounit/upholster, mořit/ebonize, leštit/polish, sklížit/glue...*

In Levin/Palmer's list of semantic classes this type of verbs belongs mostly to wide classes 9. Verbs of Putting, 45. Verbs of Change of State and 47. Verbs of Existence.

Verbs complemented by 2^{nd} level semantic role *vehicle:1* (it occurs in 153 valency frames in VerbaLex) can create following semantic groups:

modes of movement – *zrychlit/accelerate, zpomalit/slow down, brzdit/brake, couvat/back a car, zatočit/turn, předjet/overtake...*
meet with an accident – *nabourat/smash car, narazit/crash...*
transport of people – *jet/go, nastoupit/get in, vystoupit/get out, cestovat/travel, dojíždět/commute...*
transport of load – *vézt/carry, naložit/load, vyložit/unload, přepravit/transport...*
visual and acoustic signals – *houkat/hoot, troubit/toot, blikat/blink, burácet/roar...*

In Levin/Palmer's list of semantic classes this type of verbs belongs mostly to classes 11. Verbs of Sending and Carrying, 18. Verbs of Contact by Impact, 43. Verbs of Emission and 51. Verbs of Motion.

5 Conclusion

The described type of classification can be used for 2^{nd} level semantic role with reasonable frequency in VerbaLex (from 30 to 1 000 occurrences). The roles with general meaning and frequency higher than 1 000 occurrences are not suitable for this purpose (e.g. *object:1* – 2 500 occurrences). In spite of this, 2nd level semantic roles in VerbaLex present significant support for subclassification of verb semantic classes.

Acknowledgements

This work has been partly supported by the Academy of Sciences of Czech Republic under the projects 1ET100300414 and 1ET200610406 and by the Czech Science Foundation under the project 201/05/2781.

References

1. Pala, K., Ševeček, P.: Valence českých sloves. In: Proceedings of Works of Philosophical Faculty at the University of Brno, Brno, MU (1997), 41–54.
2. Žabokrtský, K.: Valency Lexicon of Czech Verbs. Ph.D. thesis, Prague (2005).
3. BalkaNet: Balkanet project website (2001–2004).
4. Fellbaum, C.e.: WordNet. An Electronic Lexical Database. MIT Press, Cambridge (1998).
5. Hlaváčková, D., Horák, A.: 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.
6. Vossen, P., Bloksma, L.e.a.: The EuroWordNet Base Concepts and Top Ontology. In: Technical Report Deliverable D017, D034, D036, WP5 EuroWordNet, LE2-4003, Amsterdam, University of Amsterdam (1998).
7. Levin, B.: English Verb Classes and Alternations: A Preliminary Investigation, Chicago, The University of Chicago Press (1993).
8. Palmer, M., Rosenzweig, J., Dang, H.T.e.a.: Investigating regular sense extensions based on intersective Levin classes. In: Coling/ACL-98, 36th Association of Computational Linguistics Conference, Montreal (1998) 293–300.