Preparing VerbaLex Printed Edition

Dana Hlaváčková, Aleš Horák, and Karel Pala

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

Abstract. In the paper, we present the current state of the development of the Czech valency dictionary called VerbaLex. It contains a list of the most frequented Czech verbs and their valency frames in the form of the complex valency frames. VerbaLex includes information about verb case and adverbial links (morphosyntactic properties) and senses captured by an inventory of two-layer semantic roles that characterize the semantics of the verb arguments.

We also present the motivation and history of the design of the complex valency frames and the VerbaLex lexicon. One of the main aims here is the support of computer analysis of Czech, thus machine-readable features of the lexicon are emphasized since the beginning. Presently, we can refer to VerbaLex electronic version with more than 10 thousand verb lemmata, as well as to its printed form with a selected subset of the most frequent verbs. The full electronic form is available on-line after registration for academic and non-commercial purposes.

Key words: verb, verb frame, verb valency, VerbaLex, WordNet, VerbNet

1 Introduction

We present to readers a new Czech valency dictionary, which is being developed in the electronic form since 2006 with the title *VerbaLex*. The dictionary contains a list of the most frequented Czech verbs and their valency frames including information about their case and adverbial links (morphosyntactic properties) and senses captured by an inventory of the complex semantic roles that characterize the semantics of the verb arguments.

The dictionary is intended for an expert public, linguists, translators, researchers in the NLP area and anybody who is interested in a deeper understanding of Czech as a mother tongue. It can also serve (as a resource) in computer applications directed to information search, summarization and possibly Machine Translation. In this context we would like to mention another valency dictionary of Czech verbs, named *Vallex*, which was prepared by the group of authors from the Institute of Formal and Applied Linguistics, Charles University [1, 6460 lex. units].

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2013, pp. 3–11, 2013. © Tribun EU 2013

A question could be raised rightly why we consider it useful to develop another valency dictionary of Czech? A following metaphor offers an answer – a language, in our case Czech, can be considered by researchers as a fortress they are trying to conquer from various angles. Thus, it is natural to approach verb valencies in Czech from different standpoints with the purpose to reach a deeper understanding of their nature. *VerbaLex* offers a different view of the Czech verb valencies than *Vallex* – the main difference consists in the conception of the semantic roles characterizing the meaning of the verb arguments (in *Vallex* named actants).

In other words, the difference lies in the approach to semantic properties of the Czech verbs – we are convinced that different solutions can be chosen and all can be reasonably justified – in practice we usually prefer the one which proves to be more fitting for the particular application. We say more about the differences between actants in *Vallex* and semantic roles in *VerbaLex* below, see especially 3.1.

2 The VerbaLex Valency Lexicon

VerbaLex is an electronic lexical database comprising verb valency frames – it has been developed in the NLP Centre, Faculty of Informatics, Masaryk University in Brno (*FI MU*) during 2006–2013. It is a result of the work, which partly belongs to the area of linguistics and partly to the field of Natural Language Processing (NLP). During the development of *VerbaLex*, we have been using various corpus resources and electronic tools, which made it possible to observe the behaviour of the verbs in their natural contexts. The main part of the database has been compiled by the annotators who relied on their linguistic competence, followed given instructions and using the accessible software tools created what is called the *basic* and *complex valency frames*. In this way the issue of the Czech verb valencies could have been captured in their complexity as much as possible.

The verb valency is understood in the database as a semantically given ability of the verb allowing it to combine with other words – the verbs are described from this point of view together with their complements both on the left and right side. Thus valency frames include two kinds of information: the *morphosyntactic* and *semantic* one. Our effort was to capture as many Czech verbs as possible, presently the *VerbaLex* comprises approximately 10 500 verb lemmata.

When compiling *VerbaLex* we have used some existing resources, in the first place the *Valenční slovník českých sloves* (*Valency Dictionary of Czech Verbs*) with working name *BRIEF* [2].

Our motivation has been an effort for a deeper understanding of the semantics of Czech verbs and their arguments and creation of the new data resources, which for Czech exist only in part. In comparison with the traditional approaches, for instance [3], we have used methods and techniques, particularly semantic networks and ontologies, which do not appear in existing Czech

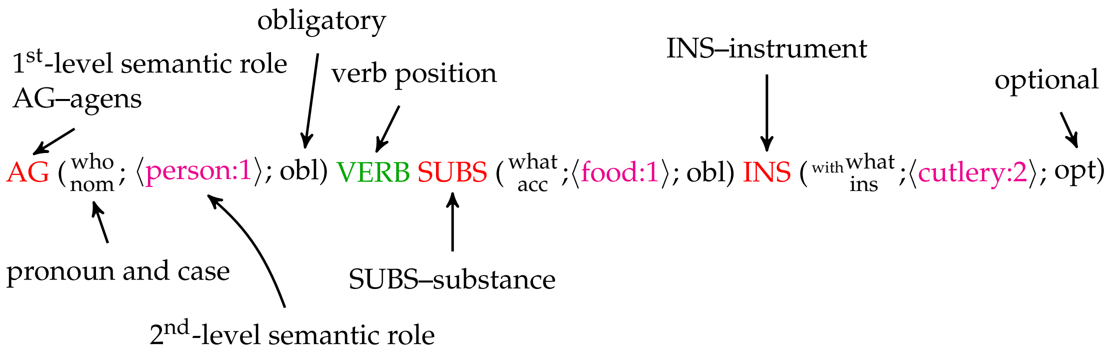

Fig. 1: Basic valency frame

grammars at all. The obtained results can be then exploited in the field of NLP, since *VerbaLex* captures semantic relations. Thus it can be used in various applications such as the intelligent search on the Web, word sense disambiguation, information extraction or text understanding [4,5].

2.1 VerbaLex Structure

After starting exploration of the suitable format for verb valencies in 2006, the structure of *VerbaLex* has consolidated in the form presented below. The database displays some basic features, in which it differs from similar dictionaries. The form of the *complex valency frames* allow us to capture the relevant information about a verb and its complements. The valency frames are assigned to individual verb senses (grouped in synonymical sets or *synsets*, see Section 2.4) and not only to individual lemmata (many synonyms share the same valency). To label the meanings of the verb complements the system of the two-level semantic roles has been used.

The basic valency frames (see Figure 1), which represent the core of *VerbaLex*, constitute the notation of the verb valency on the morphosyntactic and semantic level. The center of the frame is a marked verb position, its valency complements on the morphological level are represented by the pronominal expressions together with the respective case numbers. The notation follows the canonical word-order: “the complement on the left side – verb – the complements on the right side.” On the semantic level the verb arguments are labeled by the two-level semantic roles, which specify the semantic environment of the verb precisely. The frame contains additional information about obligatoriness and optionality of the valency complements. The basic valency frame is always related to a *subsynset*, which is a subset of the defined synonymical set.

The basic frame is a part of the complex valency frame (see Figure 2), which is always related to the one synonymical set only. Apart from the frame there is additional information which includes verb sense, aspect (see Section 2.3) and verb semantic class (see Section 3.2). For each verb its ability to form passive voice is recorded, thus it is possible to generate lists of the transitive and intransitive verbs from the database. The important feature of the Czech verbs

jíst:1 (impf), požít:2(pf), požívat:2(impf) (eat:1)
definition: *přijímat potravu* (take in solid food)
class: *eat-39.1*
passive: *yes*

jíst:1 (eat:1) ≈
-frame: **AG**($^{\text{who}}_{\text{nom}}$;⟨person:1⟩;obl) **VERB** **SUBS**($^{\text{what}}_{\text{acc}}$;⟨food:1⟩;obl)
INS($^{\text{with what}}_{\text{ins}}$;⟨cutlery:2⟩;opt)
-example: *synovec jedl zmrzlinu (impf)* (the nephew ate an ice cream)
-example: *dcera jí polévku lžící (impf)* (the daughter eats a soop with a spoon)
-use: prim
-reflexivity: no

Fig. 2: Complex valency frame

is their obligatory or optional reflexivity – we provide information about three basic types of reflexivity: proper reflexives (reflexiva tantum), i.e. verbs with obligatory reflexive particle *se*, e.g. *bát se* (*to be afraid*). Further, object reflexivity is marked when the pronoun *se* or *si* replaces object of the action (*mýt se*, *čistit si* (*wash yourself*, *clean yourself*)) as well as reciprocity, mutual activity of the two subjects (*znát se*, *milovat se* (*to know each other*, *to love each other*)). Here the verb lemma is not given with the reflexive pronoun. For the rest of the verb lemmata, we mark the fact that they have or do not have the reflexive form in the respective sense and characterize it as another (not specified) type of the reflexivity. The next relevant information is related to the behaviour of the verb in a particular context: we mark their primary (basic) usage in contrast to the figurative (metaphorical) meaning. In some cases displaying a higher frequency in corpora we also indicate the idiomatic usage.

2.2 Verb List Selection

The choice of the verb lemmata contained in the database *VerbaLex* has been based mainly on *Slovník spisovné češtiny* (Dictionary of Literary Czech, SSČ [6]) and *Slovník spisovného jazyka českého* (Dictionary of Literary Czech Language, SSJČ [7]). Moreover, the lemma selection stylistic features and frequencies of the particular verbs in the corpora *SYN2000* [8] and *ALL* (*NLP Centre FI MU*) have been considered. As a basic data resource, the *Valenční slovník českých sloves* [2] has served, which describes right side valency complements (without information about their meaning), for 15 079 Czech verbs. In *VerbaLex*, we have stored only verbs belonging to the literary Czech, where some of them can eventually have the emotional colouring. *VerbaLex* does not include verbs from colloquial Czech and dialects. We also have left aside verbs that are strongly bookish, archaic or rarely used. However, we have taken into consideration the cases when a verb is marked in a dictionary as bookish or rarely used (e.g. *pravit* (*say*, bookish form)), but they show high frequency in the corpora (verb forms like *pravit* – shows 29 397 occurrences in the corpus *ALL*), in such cases we have respected the frequencies found in corpora.

The verb lemma is always one-word, in the case of the proper reflexives (reflexiva tantum) with the reflexive particles *se*, *si*. The database does not contain negated forms of the verb lemmata, we work with the assumption that the valency frames of the negated verbs remain unchanged (except for cases with the negated genitive) and the negated forms of lemmata is possible to derive automatically. In *VerbaLex*, there is a formal way how to handle to what we call variant lemmata. In such cases the verb forms differ only in the vowel alternation, otherwise all their characteristics remain the same (e.g. *muset/musit* (*must*), *bydlit/bydlet* (*live*), *červavět/červivět* (*become wormy*)).

2.3 Verb Aspect

In the complex valency frame notation also other important information about verbs has been captured. In the first place, it is a formal notation of the verb aspect as related to the respective verb sense. The aspect identification is primarily based on the information in the dictionaries *SSČ* a *SSJČ*. If a given verb sense is valid for both its aspect forms, the verb in the first position is marked as perfective (*pf*) together with the number of the sense followed in the brackets by its imperfective form (*impf*), which automatically takes over the sense number from the perfective and it is not necessary to indicate it again. The verbs with two aspects are denoted as biaspectual *biasp*. Iterative verb forms are not stored in the database. Their forms are derived very regularly and can be automatically added from the existing morphological database to the *VerbaLex* at any time.

If a verb sense is valid for just one aspect form or a verb is one-aspectual, we mark the perfective or imperfective verb form independently with its own sense number. In the case of the perfective verbs derived by prefixation from the imperfective verb, we consider as aspect pairs only the cases, which are explicitly marked in the dictionaries *SSČ* a *SSJČ* (prefixation with the aspect prefix only, e.g. *uvařit* (*to cook*), *učesat* (*to comb*)). Other prefixed verbs are marked either independently or in the pair with the respective secondary imperfective.

2.4 Synonymy and Verb Senses

As we have mentioned above, verbs in *VerbaLex* are organized in synonymical sets (*synsets*). In synsets, each verb lemma (and its variants) are marked with ordinal number denoting their sense.[1] This denotation directly corresponds to the numbering in the Czech WordNet, see Section 3. The appropriate synonyms have been chosen and verified in *Slovník českých synonym* (Dictionary of Czech Synonyms, SČS [9]). Each synset is accompanied with a short definition of its meaning. The definitions (with necessary modifications) are formulated on the basis of the lexicographic definitions in *SSČ* and *SSJČ*. In the specification

[1] The tuple "lemma:sense" is often called a *literal*.

of particular verb senses, we often cannot always use SSČ and SSJČ directly. Their approach differs in many cases and they state different number of verb senses. These dictionaries are also not sufficiently up-to-date source of current language. In case of a verb sense, which was not found in the dictionaries, the verb occurrences in new contexts were verified on corpus data (*SYN2000*, *ALL*). If the number of the verb sense occurrences reached adequate frequency, the sense was added to *VerbaLex* with new sense number. Obsolete and rare cases from the dictionaries are not used in *VerbaLex* at all. In accordance with the *WordNet* structure,[2] the verb sense determination is often more fine-grained than in usual Czech dictionaries.

3 Semantic Description Layer – VerbaLex and WordNet

One of the main differences of *VerbaLex* when compared to *Vallex* is the narrow connection of *VerbaLex* to the *WordNet* semantic network (*Princeton WordNet*, *PWN* [10]) since the very beginning. During the *BalkaNet* EU project in 2002, the Czech WordNet (*CzWn*) structure was supplemented with basic valency frames including semantic roles. According to the *PWN* structure, the frames were linked to whole synsets instead of individual verb lemmata. For the same reason, the frames were divided according to particular verb senses. This approach was then adopted in the *VerbaLex* preparation procedures.

The verb synonymy is here understood in a broader sense than usual. The synset participants are often *near synonyms*, which cannot be freely interchanged in the same contexts. Synsets in *PWN* are interlinked by several kinds of relations, the most important being the hypernym/hyponym hierarchy. The hypernym/hyponym relations are most significant for nouns, in case of verbs this relation corresponds to *troponymy*, the relation of doing something in a specific manner.[3]

In the late 90s, the *PWN* approach was applied within the EU projects *EuroWordNet-1* and *2* (*EWN*), in which new national *WordNets* were created for Dutch, Italian, Spanish, French, German, Czech and Estonian. The synsets in the national *WordNets* were interlinked by means of *Interlingual Index* (*ILI*) describing the translational equivalents. In each language, for which a *WordNet* was created, we can find at least 15 000 synsets with equivalents in *PWN*.

In *VerbaLex*, all verb senses are directly linked to their English equivalents in *PWN*. The newly added synsets were linked to the *PWN* English synsets using the WordNet Assistant tool [11]. Appropriate equivalents could be found for 85 % out of 3 686 new synsets. In 15 % there is usually not direct lexicalized equivalent – for perfective, reflexive or prefixed verbs, or verbs with expressive or metaphoric meaning. For example, the Czech verb *povyskočit* ("jump up a little") is not found in any standard bilingual dictionary. The same holds for *povyskakovat* ("jump out of [something] one after another") and many other

[2] see Section 3

[3] E.g. "pohybovat se" (to move) → "chodit" (to walk) → "klopýtat" (to stumble).

Substance – in VerbaLex a semantic role:
1st-level – **SUBS**
2nd-level, PWN hypernym – **substance:1**
Two-layer semantic role – **SUBS(substance:1)**
Hyponymic lexical units as specifiers:
SUBS⟨solid:1⟩, *SUBS⟨liquid:3⟩*, *SUBS⟨gas:2⟩*, *SUBS⟨food:1⟩*, *SUBS⟨beverage:1⟩*, ...
Hyponymic subclass of particular examples:
SUBS⟨beverage:1⟩ = milk:1, alcohol:1, chocolate:1, fruit juice:1, soft drink:1, coffee:1, tea:1, drinking water:1, ...

Fig. 3: Example of a two-layer semantic role

verbs with two prefixes, e.g. *povyřizovat* (“finish doing things successively”), *dovyplnit* (“fill in an extra information”).

3.1 Semantic Roles

Within the *EWN* projects, the core of the shared interlingual lexicon was defined as formed by the *Top Ontology* and a larger set of *Base Concepts*.[4] The top ontology also inspired the *VerbaLex* system of two-level semantic roles. Above all, we have selected the concepts covering large classes of lexical meanings. The classes correspond to the top hypernyms in the *PWN* hierarchy. We have chosen the hypernyms that best reflect the relevant meanings of the semantic roles and that are branching to expected hyponyms. *VerbaLex* 1st-level semantic roles use literals with sense number 1 or 2, i.e. basic meanings, which belong to the set of base concepts. The whole set of 1st-level roles is currently formed by 32 semantic roles, which describe very general meanings reflecting the reality. Each role covers one well recognized and specified meaning area, e.g. *ARTifact*, *ACTivity*, *INStrument*, *COMmunication*, *EVENt*, *LOCation*, or *TIME*.

The 2nd-level roles use direct hyponyms from *PWN* serving as a specification of the “most expected” meaning of this verb argument. The hyponyms of such literals can then serve as instances of the appropriate class. An example can be two-level roles denoting all substances (solid, liquid or gas), see Figure 3. The usage of 2nd-level roles can also be understood as subcategorization features, or selectional restrictions. They form an open system of labels, which can be continuously extended with regard to current applications.[5] The motivation for such approach lies in the aspiration to obtain a detailed description of particular verb senses. In this respect, *VerbaLex* fundamentally differs from the *Vallex* lexicon.

[4] At the beginning the set included about 1 000 base concepts, which was later extended to 8 000 concepts.

[5] *VerbaLex* currently contains 811 2nd-level semantic roles.

3.2 Semantic Classes of Verbs

The *VerbaLex* database describes not only meanings of the verb arguments, but also the meaning of the verb itself, which are one of the principal factors of its valency frames at both syntactic and semantic levels. The verb meaning is, besides the human readable definition, captured by detailed classification using verb semantic classes. Experimentally, we have chosen the classification system of English verbs by B. Levin [12], which builds upon the syntactic and semantic features of English verbs. The system divides verbs according to alternations of their participants. Within the *VerbNet* project of M. Palmer 48 basic semantic classes of Levin were extended to 83 classes (numbered 9.–91. [13]). Ambiguous verbs, originally instantiated in multiple classes, were detached to individual classes with their own meaning.

In *VerbaLex*, we have adapted the original set of semantic classes from *VerbNet* (numbered from 9 according to Levin and Palmer) and divided some of them to meaning subclasses resulting in 109 current verb class repository. The classes are originally based on the description of changes in the argument structure of English verbs, but after the adaptation they serve the purpose very well also for Czech.

4 Conclusions

In this paper, we have presented in detail the current state of the development of the *VerbaLex* valency lexicon of Czech verbs, including the electronic version with 10 449 verb lemmata, as well as its printed form with selected subset of the most frequent verbs. The full electronic form is available on-line after registration for academic and non-commercial purposes. The electronic database definitely offers more information by direct connection to the Czech and English WordNet and the possibility of intelligent browsing and searching. Even though the printed form of *VerbaLex* is a limited volume, we believe that a handy book also has its advantages.

Acknowledgements This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013 and by the Ministry of the Interior of CR within the project VF20102014003.

References

1. Lopatková, M., Žabokrtský, Z.: Vallex (2008) `http://ucnk.ff.cuni.cz`, 6460 lex.units.
2. Pala, K., Ševeček, P.: Valence českých sloves (Valencies of Czech Verbs). In: Sborník prací Filosofické fakulty Masarykovy university, A45, Brno (1997) 41–54
3. Svozilová, N., Prouzová, H., Jirsová, A.: Slovesa pro praxi: valenční slovník nejčastějších českých sloves (Verbs for Practice: Valency Dictionary of the Most Frequent Czech Verbs). Academia, Prague (1997)

4. Jakubíček, M., Horák, A.: Punctuation Detection with Full Syntactic Parsing. Research in Computing Science, Special issue: Natural Language Processing and its Applications **46** (2010) 335–343
5. Hlaváčková, D., Horák, A., Kadlec, V.: Exploitation of the VerbaLex Verb Valency Lexicon in the Syntactic Analysis of Czech. In: Proceedings of Text, Speech and Dialogue 2006, Brno, Czech Republic, Springer-Verlag (2006) 79–85
6. Filipec, J., Daneš, F., Mejstřík, V., eds.: Slovník spisovné češtiny pro školu a veřejnost (Dictionary of Literary Czech for School and Public). Academia, Prague (2000)
7. Havránek, B., ed.: Slovník spisovného jazyka českého (Dictionary of Literary Czech Language). Academia, Prague (1989)
8. Intitute of Czech National Corpus, FA CU: Czech National Corpus – SYN2000 (2000) `http://ucnk.ff.cuni.cz`.
9. Pala, K., Všianský, J.: Slovník českých synonym (Dictionary of Czech Synonyms). Lidové noviny, Prague (1996)
10. Fellbaum, C., ed.: WordNet: An Electronic Lexical Database. Language, Speech, and Communication. MIT Press, Cambrige (1998)
11. Němčík, V., Pala, K., Hlaváčková, D.: Semi-automatic linking of new czech synsets using princeton wordnet. In: Proceedings of the Intelligent Information Systems XVI Conference (IIS'08), Warszawa, Academic Publishing House EXIT (2008) 369–374
12. Levin, B.: English Verb Classes and Alternations: A Preliminary Investigation. University of Chicago Press, Chicago (1993)
13. Palmer, M., Rosenzweig, J., Dang, H.T., Kipper, K.: Investigating regular sense extensions based on intersective levin classes. In: Proceedings of the 17th international conference on Computational linguistics, Association for Computational Linguistics (1998) 293–299