Transformation of WordNet Czech Valency Frames into Augmented VALLEX-1.0 Format

Dana Hlaváčková, Aleš Horák

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

Abstract

The paper presents details and comparison of two valuable language resources for Czech, two independent verb valency frames electronic dictionaries. The FIMU verb valency frames dictionary was designed during the EuroWordNet project and contains semantic roles and links to the Czech wordnet semantic network. The VALLEX 1.0 format is based on the formalism of the Functional Generative Description (FGD) and was developed during the Prague Dependency Treebank (PDT) project. We present the tools and approaches that were used within the process of adopting the FIMU Vallex format for the wordnet enriched valency frames.

1. Introduction

The beginnings of building the verb valency frame dictionary at the Faculty of Informatics at Masaryk University (FIMU) dates back to 1997 (Pala and Sevecek, 1997). Since then, the dictionary has undergone a long development and has been used in various tools from semantic classification to syntactic analysis of Czech sentence (Smrz and Horak, 1998). Currently, the dictionary plays a key role within an experimental high-coverage syntactic analysis using data from the Czech WordNet.

The FIMU dictionary is being actively developed, checked and supplemented with new data. The last stage of the list of valency frames was originally started during the creation of Czech WordNet within the EuroWordNet project (Vossen, 1998). 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 (Stranakova-Lopatkova and Zabokrtsky, 2002). Currently, the FIMU dictionary contains 3249 verbs (being more than twice the number of verbs in VALLEX 1.0) which, when gathered in synonymic groups, share 1646 verb frames.

2. Linguistic requirements for the FIMU Vallex format

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

The lexical units in WordNet are organized into synsets (set of synonyms) arranged in the hierarchy of word meanings (hyper-hyponymic relations). FIMU Vallex differs from VALLEX 1.0 in augmentation of original format, detailed differentiation of valency frames and above all semantic roles (deep cases). For that reason the word entries in FIMU Vallex are lemmata with synonymic relation (synset subsets) followed by sense number (standard Princeton WordNet notation) and with common valency frame. The standard definition of synonymity says that two synonymic words can be allways 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 word entry is one lemma, possibly two or more lemmata in case of lemma variants (the lemmata with small phoneme alternation that are interchangeable in any context without any change of the meaning – `myslet/myslit` (to think)). Lemma variants in FIMU Vallex are considered as independent lemmata and they are distinguished by sense number from WordNet. An example of two verb frame entries in VALLEX 1.0 and FIMU Vallex is displayed in the Figure 1.

In a free-word order language like Czech the position of the verb within the verb frame is usually not strictly specified. However, there are some cases, where the verb frame has to obey certain rules – e.g. sentence *Dalo se do deste* (It started to rain) cannot contain any subject. Such requirements are captured within the FIMU Vallex with new `VERB` semantic role, which marks the (usual) position of the verb in its verb frame. Such default verb position is also very useful for the sake of generation of Czech sentences as an output of a question-answering machine.

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

Princeton WordNet – awaken:1, wake:5,
waken:1, rouse:4, wake up:1, arouse:5
Definition: cause to become awake or
conscious
FIMU Vallex: budit:1 impf. / vzbudit:1 pf. /
probudit:1 pf.
frame: CAUSE <cause:4>$^{\text{obl}}_{\text{co1}}$ VERB
PAT <person:1>$^{\text{obl}}_{\text{koho4}}$
example: probudil me hluk pf. / that noise
awoke me
example: budi me budik impf./ an alarm clock
wakes me up

There are some differences in respective frame entries as well. The constituent elements of frame entries are en-

Lemma variants:

Princeton WordNet – think about:1
Definition: have on one's mind, think about actively
VALLEX 1.0: myslet$_3$ / myslit$_3$
FIMU Vallex: myslet:3 / myslit:3 / pamatovat:2

Word entries:

Princeton WordNet – carry through:1, accomplish:1, execute:3, carry out:1, action:2, fulfill:1, fulfil:1
Definition: put in effect
VALLEX 1.0: dokoncit$_1$
FIMU Vallex: splnit:1 / dokoncit:2 / vykonat:4 / dorazit:5

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

riched by pronominal terms with number of morphological case, which allow to differentiate animate or inaminate agent position for example.

Princeton WordNet – expose:3, exhibit:2, display:1
Definition: to show, make visible or apparent
VALLEX 1.0: vystavit$_1$
frame: ACT$_1^{obl}$ PAT$_4^{obl}$ LOC^{typ}
FIMU Vallex: vystavit:1
frame: AG <person:1>$^{obl}_{kdo1}$ VERB
ART <creation:2>$^{obl}_{co4}$
example: the young artist exhibits his first paintings
frame: AG <institution:1>$^{obl}_{co1}$ VERB
ART <creation:2>$^{obl}_{co4}$
example: the Metropolitan Museum exhibits Goya's works

The main difference between VALLEX 1.0 and FIMU Vallex valency frames is evident from the stock list of semantic roles (functors in Vallex) and from the way of its notation. The functors used in VALLEX 1.0 valency frames seem to be too general and that does 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 (Vossen et al., 1998). On the second level, we use some 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 `beverage:1` is hypernym for any liquid suitable for drinking.

Princeton WordNet – drink:1, imbibe:3
Definition: take in liquids
FIMU Vallex: pit:1
frame: AG <person:1,animal:1>$^{obl}_{kdo1}$ VERB
SUBS<beverage:1>$^{obl}_{co4}$
example: my brother drinks beer, horse drinks water

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 completely 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 – curl:4, wave:4
Definition: twist or roll into coils or ringlets
VALLEX 1.0: natočit$_4$
frame: ACT$_1^{obl}$ PAT$_4^{obl}$ BEN$_3^{typ}$
FIMU Vallex: natočit:3
frame: AG <person:1>$^{obl}_{kdo1}$ VERB
PART <hair:6>$^{obl}_{co4}$ PAT <person:1>$^{obl}_{komu3}$
example: Mary curls her friend's hair

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`, `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 FIMU Vallex semantic roles that are used in the presented examples is displayed in the Table 1.

The agent position in valency frame is understood as 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

Table 1: List of semantic roles from FIMU Vallex that are used in examples.

AG	the semantic role of the animate entity that instigates or causes the hapening 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
CAUSE	any entity that causes events to happen
OBJ	a tangible and visible entity; an entity that can cast a shadow
INFO	a message received and understood that reduces the recipient's uncertainty

FIMU Vallex are able to define cause of action quite precisely. The main semantic role AG is completed by 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), 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 – give birth:1, deliver:12, bear:2, birth:1, have:18
> Definition: give birth (to a newborn)
> FIMU Vallex: plodit:1 / rodit:1 / mit:4
> frame: AG $<\text{woman:1}>^{\text{obl}}_{\text{kdo1}}$ VERB
> PAT $<\text{child:2}>^{\text{obl}}_{\text{koho4}}$
> example: my wife gave birth to our son

Every valency frame starts always with functor ACT in VALLEX 1.0. In our opinion, it is useful to differentiate the sense of 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 occupied by another semantic roles too. The left position from the verb 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 – run out:1
> Definition: become used up; be exhausted
> FIMU Vallex: dojit:8 / spotrebovat se:1 / vycerpat se:1
> / vydat se:2
> frame: OBJ $<\text{object:1}>^{\text{obl}}_{\text{co1}}$ VERB
> example: our supplies finally ran out

> Princeton WordNet – transpire:3
> Definition: come to light; become known
> FIMU Vallex: prozradit se:1 / ukazat se:3
> frame: INFO $<\text{fact:1}>^{\text{obl}}_{\text{co1}}$ VERB
> example: the secret transpired

3. Implementation of editing and exporting tools

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

The main interactive tool for user editing of the valency dictionary, `vallex.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 `vallex.sh` currently offers these functions to the editing user:

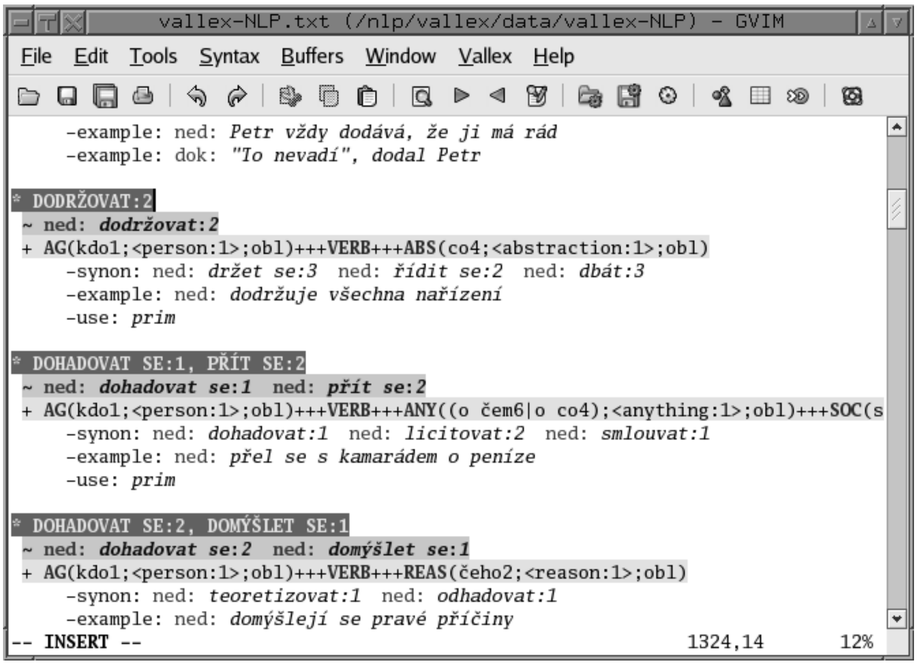
- free editing of the dictionary entries
- regular expression searching in the dictionary
- template-based adding of a new verb entry or a new verb frame to the current entry
- menu-based adding of new semantic role to the current frame
- multilevel folding – hiding/unhiding of valency attributes, valencies or full valency frames
- visual marking the current frame for further inquiry

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 FIMU Vallex. 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.



```
    -example: ned: Petr vždy dodává, že ji má rád
    -example: dok: "To nevadí", dodal Petr

* DODRŽOVAT:2
 ~ ned: dodržovat:2
 + AG(kdo1;<person:1>;obl)+++VERB+++ABS(co4;<abstraction:1>;obl)
    -synon: ned: držet se:3  ned: řídit se:2  ned: dbát:3
    -example: ned: dodržuje všechna nařízení
    -use: prim

* DOHADOVAT SE:1, PŘÍT SE:2
 ~ ned: dohadovat se:1  ned: přít se:2
 + AG(kdo1;<person:1>;obl)+++VERB+++ANY((o čem6|o co4);<anything:1>;obl)+++SOC(s
    -synon: ned: dohadovat:1  ned: licitovat:2  ned: smlouvat:1
    -example: ned: přel se s kamarádem o peníze
    -use: prim

* DOHADOVAT SE:2, DOMÝŠLET SE:1
 ~ ned: dohadovat se:2  ned: domýšlet se:1
 + AG(kdo1;<person:1>;obl)+++VERB+++REAS(čeho2;<reason:1>;obl)
    -synon: ned: teoretizovat:1  ned: odhadovat:1
    -example: ned: domýšlejí se pravé příčiny
-- INSERT --                                              1324,14       12%
```

Figure 2: The tool for editing verb valency frames dictionary in the FIMU Vallex format.

```
<headword_variants>
 <headword_lemma ord='1' aspect='pf' sense='1' asp_counterpart='2'>ocenit</headword_lemma>
 <headword_lemma ord='2' aspect='impf' asp_counterpart='1'>ocenovat</headword_lemma>
 <headword_lemma ord='3' aspect='impf' sense='1'>cenit</headword_lemma>
 <headword_lemma ord='4' aspect='pf' sense='1' asp_counterpart='5'>ohodnotit</headword_lemma>
 <headword_lemma ord='5' aspect='impf' asp_counterpart='4'>hodnotit</headword_lemma>
</headword_variants>
```

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

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.

4. Conclusions and Future Directions

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

The nearest development of the FIMU valency dictionary includes 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 analyser.

5. 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.

6. References

Pala, Karel and Pavel Sevecek, 1997. Valence ceskych sloves (valencies of Czech verbs). In *Proceedings of Works of Philosophical Faculty at the University of Brno*. Brno: Masaryk University.

Smrz, P. and A. Horak, 1998. Determining type of TIL construction with verb valency analyser. In *Proceedings of SOFSEM'98*. Berlin: Springer-Verlag.

Stranakova-Lopatkova, M. and Z. Zabokrtsky, 2002. Valency dictionary of czech verbs: Complex tectogrammatical annotation. In C. Paz Surez Araujo M. Gonzlez Rodrguez (ed.), *LREC2002, Proceedings*, volume III. ELRA.

Vossen, P., L. Bloksma, et al., 1998. The EuroWordNet base concepts and top ontology. Technical Report Deliverable D017, D034, D036, WP5 EuroWordNet, LE2-4003, University of Amsterdam.

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