Mapping Semantic Information from FrameNet onto VALLEX[1]

Václava Kettnerová and Markéta Lopatková
Charles University in Prague, Institute of Formal and Applied Linguistics
{kettnerova,lopatkova}@ufal.mff.cuni.cz

1. Information on syntactic and semantic properties of verbs, which are traditionally considered as a center of sentence, plays a key role in many rule-based NLP tasks as information retrieval, text summarizing, question answering, machine translation, etc. Lexical resources providing such information are designed, however, within different theoretical frameworks whose theoretical assumptions are reflected in annotation schemes. As a result, there are great differences between individual lexical resources. However, the different theoretical background taken in individual lexical resources has another consequence: each lexical resource captures different types of information. Use of information from several lexical resources then represents an effective way of enriching a particular lexical resource.

On the other hand, differences in theoretical assumptions taken in lexical resources bring several difficulties with mapping information: the different level of granularity in word sense disambiguation represents a typical example. Moreover, other requirements for harmonizing linguistic information are imposed on interlinking information from different-lingual lexical resources: a fundamental prerequisite for successful mapping represents especially an accurate translation.

In this contribution, we introduce a project aimed at enhancing a valency lexicon of Czech verbs, VALLEX, with semantic information from FrameNet. First, we classify verbs from chosen groups of verbs, namely verbs of communication, mental action, exchange, motion, transport and psych verbs, into more coherent semantic classes based on semantic frames from FrameNet. Second, we assign frame elements as semantic roles to each valency complementation of given verbs.

This project represents an example of mapping information from different-lingual lexical resources, FrameNet and VALLEX. These resources are based on different theoretical assumptions: VALLEX takes primarily syntactic criteria in describing valency whereas FrameNet adopts more semantically oriented approach to valency. Furthermore, we have to cope with the difficulty with the different level of granularity in word sense disambiguation made in VALLEX and FrameNet.

2. Let us briefly introduce these two lexical resources. **VALLEX 2.5** (http://ufal.mff.cuni.cz/vallex/2.5/) provides information on the valency structure of verbs in their particular senses: on the number of valency complementations, on their type labeled by functors, and on their morphemic forms, (Žabokrtský, Lopatková, 2007). VALLEX 2.5 describes 2730 verb lexemes containing about 6460 lexical units (LUs in the sequel) typically corresponding to one sense. At present, more than 44% of LUs are divided into heterogeneous 'supergroups' − as 'communication', 'mental action', 'motion', 'exchange', 'transport', etc. −, which represent rather tentative classification, based primarily on similar morphosyntactic

[1] This research was carried under the MŠMT ČR project No. MSM0021620838 and partially under GA UK grants No. 7982/2007 and 4200/2009.

patterns (number of valency complementations, their morphemic forms and (for some groups) specific syntactic properties) and similar semantics.

Key information on valency is stored in valency frame. VALLEX 2.5, which is closely related to the Prague Dependency Treebank 2.0, (Hajič et al., 2006) takes the *Functional Generative Description* (FGD in the sequel) as its theoretical background, (Sgall, et al., 1986). FGD applies more syntactically oriented approach to valency, (Panevová, 1994). Valency complementations are sorted out into inner participants (arguments in the sequel) and free modifications (adjuncts in the sequel). Both arguments and adjuncts may be obligatory or optional. Five verbal arguments are determined rather on the basis of syntactic behavior of verbs: 'Actor' (labeled ACT), 'Patient' (PAT), 'Effect' (EFF), 'Addressee' (ADDR) and 'Origin' (ORIG). In contrast to arguments, adjuncts are semantically distinctive.

FrameNet (http://framenet.icsi.berkeley.edu/) is an on-line lexical database documenting semantic and syntactic combinatory possibilities (valences) of each word in each of its senses, (Baker et al., 1998). FrameNet contains more than 10 thousand LUs (pairs consisting of a word and its meaning) in more than 825 semantic frames (SFs in the sequel), exemplified by more than 135 thousand annotated sentences. Each LU evokes a particular SF underlying its meaning. Each SF is conceived as a "conceptual structure describing a particular type of situation, object, or event", (Ruppenhofer et al., 2006). Each SF contains the so-called frame elements (FEs in the sequel), i.e., semantic participants of such situations.

We focus on enhancing VALLEX with missing semantic information, namely semantic classes and semantic roles. Classifying LUs into semantic classes enables us to observe relation between semantic properties of LUs and their syntactic behavior. Furthermore, semantic roles allow us to draw inferences on lexical entailment imposed by LUs on their valency complementations. For illustration, LUs described by the same valency frame remain indistinct in VALLEX, despite being semantically different, see pairs of sentences (1)-(2) and (3)-(4). Mapping semantic information from FrameNet onto these pairs allows us to differentiate between the given LUs: *vymyslet* 'to think' is classified as belonging to 'Invention' and its valency complementations 'Actor' and 'Patient' are mapped onto FEs 'Cognizer' and 'Invention', respectively (example 1), whereas SF 'Self_motion' is assigned to *vyjít* 'to climb' and FEs 'Self_mover' and 'Path' to 'Actor' and 'Patient', respectively (example 2). Similarly, SF 'Telling' and 'Bringing' correspond to LUs from examples (3) and (4), respectively. Then their arguments 'Actor', 'Addressee' and 'Patient' are described by FEs 'Speaker', 'Addressee' and 'Message' in case of *vyprávět* 'to tell', and by FEs 'Agent', 'Goal' and 'Theme' in case of *přinést* 'to bring', respectively.

(1) *Radní*.ACT *vymysleli nový plán*.PAT *rozvoje města.*
Eng. Councilmen.ACT thought a new plan.PAT for development of the city.
(2) *Turisté*.ACT *vyšli kopec*.PAT
Eng. The tourists.ACT climbed the hill.PAT

(3) *Matka.*ACT *vyprávěla dětem.*ADDR *pohádku.*PAT
Eng. The mother.ACT told the children.ADDR the fairy-tale.PAT
(4) *Jana*.ACT *přinesla otci.*ADDR *dárek.*PAT
Eng. Jane.ACT brought the father.ADDR a gift.PAT

3. In the project, we translated each LU belonging to groups of verbs 'communication', 'mental action', 'psych verbs', 'exchange', 'motion', 'transport' from Czech into English, (1.320 verbs in total). Then the human annotators had to indicate an appropriate SF (unambiguous assignment of SF) or more than one SF (ambiguous assignment of SF) for these LUs in FrameNet. The annotators could also conclude that no SF corresponds to a given Czech LU.

If an appropriate SF was indicated, then FEs corresponding to this SF were mapped onto valency complementation(s) of the given Czech LU. The feasibility of this task was proven by the achieved inter-annotator agreement (IAA) measured on the groups of verbs of communication and exchange (Kettnerová, et al., 2008a; Kettnerová, et al., 2008b):

- assigning SFs: IAA 85.9% for verbs of communication and 78.5% for verbs of exchange (κ statistics 0.82 and 0.73, respectively) and
- assigning FEs: IAA 95.6% for verbs of communication and 91.2% for verbs of exchange (κ statistics 0.95 and 0.91, respectively)

The most frequently assigned SFs include:

- 'Statement', 'Request', and 'Telling' (for verbs of communication),
- 'Coming_to_believe', 'Becoming_aware', and Cogitation' (for verbs of mental action),
- 'Experiencer_obj', 'Cause_to_experience', and 'Experiencer_subj' (for psych verbs),
- 'Giving', 'Getting', and 'Exchange' (for verbs of exchange),
- 'Self_motion', 'Motion', and 'Arriving' (for verbs of motion) and
- 'Cause_motion', 'Bringing', and 'Removing' (for verbs of transport).

4. Finally, we proposed a method of enhancing the valency lexicon with semantic classes and semantic roles. In classifying Czech LUs and assigning semantic roles to their valency complementations, the semantic relation of 'Inheritance' plays a key role. This relation links such SFs which share basic semantic properties. Therefore, each child frame inherits semantics from its parent frame(s). As semantic classes, appropriate upper level SFs from this relation are chosen (top level SFs – represented by non-lexical and abstract SFs or SFs indicating a very general event – were omitted); i.e., each Czech LU was classified according to the ancestor SF. This method allows us to overcome the problem with coarser level of granularity made in VALLEX.

Let us demonstrate the principles of this classification on the verb *vyhnout se*$_{pf}$*, vyhýbat se*$_{impf}$ 'to sidestep'. This verb belongs to SF 'Dodging' whose upper level ancestor SF in the relation of 'Inheritance' is represented by the SF 'Avoiding'. Thus the verb *vyhnout se*$_{pf}$*, vyhýbat se*$_{impf}$ 'to sidestep' is included in the semantic class 'Avoiding'. The same class is assigned also to verbs belonging to other descendant SF of 'Avoiding', namely 'Evading' (e.g., *uhnout*$_{pf}$*, uhýbat*$_{impf}$ 'to dodge', *utéci/utéct*$_{pf}$*, utíkat*$_{impf}$ 'to flee', *uniknout*$_{pf}$, *unikat*$_{impf}$ 'to elude', *ujet*$_{pf}$*, ujíždět*$_{impf}$ 'to get away').

However, in case that Czech LU exhibits different morphosyntactic properties than LUs assigned by ancestor SF, we exploit SF from the lower level of the relation of 'Inheritance'. E.g., the verb *doprovodit*$_{pf}$*, doprovázet*$_{impf}$ 'to accompany', belongs to SF 'Cotheme' with the ancestor SF 'Self_motion'. Since this verb has different valency frame in Czech (obligatory 'Patient') than verbs onto which the SF 'Self_motion' was mapped (e.g., *běhat*$_{impf}$ 'to run', *kráčet*$_{impf}$ 'to march', *létat*$_{impf}$ 'to fly'), the SF 'Cotheme' from the lower level of the relation of 'Inheritance' was exploited.

We set 70 SFs in total as candidates for semantic classes for verbs from the above mentioned 6 groups of verbs:

- Communication (9 classes, 68% of verbs of communication were classified into these semantic classes)
- Mental action (29 classes, coverage 58% of verbs of mental action)
- Psych verbs (2 classes, coverage 13.5%)
- Exchange (10 classes, coverage 98%)

- Motion (12 classes, coverage 72.3%
- Transport (8 classes, coverage 76.5%)

5. Based on SFs mapping, we enhanced the valency lexicon with semantic roles. For this purpose, we exploit FEs from the ancestor SFs of the relation of 'Inheritance' that were chosen as semantic classes. For illustration, as semantic roles, FEs 'Agent', 'Undesirable_situation', and the others were mapped on the valency complementations of the verb *vyhnout se*$_{pf}$, *vyhýbat se*$_{impf}$ 'to sidestep', included in the semantic class 'Avoiding'. We obtained 282 FEs in total as candidates for semantic roles for the mentioned 6 'supergroups' of verbs (only core FEs as the most important ones are counted). The coverage for particular groups follows:

- almost 53% of valency complementations of verbs of communication,
- 38.5% of valency complementations of verbs of mental action,
- 12.4% of valency complementations of psych verbs,
- 95.4% of valency complementations of verbs of exchange,
- 69.2% valency complementations of verbs of motion, and
- 91.8% valency complementations of verbs of transport.

The differences in coverage are given by the different coverage of the relation of 'Inheritance' in FrameNet.

6. In conclusion, we introduce the project aimed at enhancing the valency lexicon with missing semantic information – semantic classes and semantic roles. For this purpose, we made use of FrameNet data. We proposed a method of overcoming the problem with finer granularity of word sense disambiguation made in FrameNet. This method is based on the relation of 'Inheritance'. As a result, the 6 'supergroups' of verbs were classified into more coherent semantic classes and semantic roles were assigned to their valency complementations. As to future work, we intend to experiment with other groups of verbs and to increase coverage of semantic information following the progress made in FrameNet.

References:

Baker, C.F. – Fillmore, C.J. – Lowe, J.B. (1998): The Berkeley FrameNet project. In *Proceedings of the COLING-ACL*. Montreal, Canada.

Hajič, J. et al. (2006): *Prague Dependency Treebank 2.0*. Philadelphia, PA, USA, Linguistic Data Consortium.

Kettnerová, V. – Lopatková, M. – Hrstková, K. (2008a): Semantic Classes in Czech Valency Lexicon: Verbs of Communication and Verbs of Exchange. In Proceedings of TSD 2008, LNAI 5246, Berlin Heidelberg, Springer-Verlag, pp. 109-116

Kettnerová, V. – Lopatková, M. – Hrstková, K. (2008b): Semantic Roles in Valency Lexicon of Czech Verbs: Verbs of Communication and Exchange. In Advances in Natural Language Processing; Proceedings of GoTAL 2008, LNAI 5221, Berlin Heidelberg, Springer-Verlag, pp. 217-221

Panevová, J. (1994): Valency Frames and Meaning of the Sentence. In Luelsdorff, P.A. (ed.): *The Prague School of Structural and Functional Linguistics*. Amsterdam, Philadelphia, John Benjamin Publishing Company, pp. 223−243.

Ruppenhofer, J. – Ellsworth, M. – Petruck, M.R.L. – Johnson, C. – Schefczyk, J. (2006): *FrameNet II: Extended Theory and Practice*.

http://framenet.icsi.berkeley.edu/book/book.html/

Sgall, P. – Hajičová, E. – Panevová, J. (1986): *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht, Reidel.

Žabokrtský, Z. – Lopatková, M. (2007): Valency Information in VALLEX 2.0: Logical Structure of the Lexicon. *The Prague Bulletin of Mathematical Linguistics* 87, pp. 41-60.