Semantic Classes in Czech Valency Lexicon: Verbs of Communication and Verbs of Exchange*

Václava Kettnerová, Markéta Lopatková and Klára Hrstková

Charles University in Prague, MFF ÚFAL, Prague
{kettnerova,lopatkova,hrstkova}@ufal.mff.cuni.cz

Abstract. We introduce a project aimed at enhancing a valency lexicon of Czech verbs with coherent semantic classes. For this purpose, we make use of FrameNet, a semantically oriented lexical resource. At the present stage, semantic frames from FrameNet have been mapped to two groups of verbs with divergent semantic and morphosyntactic properties, verbs of communication and verbs of exchange. The feasibility of this task has been proven by the achieved inter-annotator agreement – 85.9% for the verbs of communication and 78.5% for the verbs of exchange. As a result of our experiment, the verbs of communication have been classified into nine semantic classes and the verbs of exchange into ten classes, based on upper level semantic frames from FrameNet.

1 Introduction

Information on syntactic and semantic properties of verbs is crucial for a wide range of computational tasks. Lexical resources containing such information have been created with different theoretical backgrounds and with different goals. As a result, they store different information. Combining these resources is an effective way to gain more data for NLP tasks.

In this paper, we report on introducing semantic classes to VALLEX [1] while making use of FrameNet data. Our motivation has (i) a practical aspect – to provide available data for NLP tasks, such as generation, question answering, or information retrieval, and (ii) a theoretical aspect – semantic classes enable us to generalize about relations between the semantics of verbs and their syntactic behavior.

As a first step, we experimented with two groups of verbs with divergent semantic and morphosyntactic properties, verbs of communication and verbs of exchange. First, semantic frames from FrameNet were manually assigned to these verbs. Then the hierarchical network of relations in FrameNet between the semantic frames was used for sorting the Czech verbs into coherent semantic classes. Manual annotation is highly time consuming, however, it allows us to reach the desired quality.

* The research reported in this paper is carried under the project of the Ministry of Education, Youth and Sports No. MSM0021620838 (Objects of Research), under the grants LC536 (Center for Computational Linguistics II) and GA UK 7982/2007.

The present paper is structured as follows: In Section 2, we introduce VALLEX and FrameNet and the motivation for our experiment. Section 3 describes our experiment with mapping the FrameNet semantic frames to the two groups of Czech verbs in VALLEX. Section 4 provides an analysis of the obtained data. Section 5 presents a method of classifying verbs in VALLEX while making use of upper level semantic frames in FrameNet. Finally, results and open questions are summarized.

2 Two Lexical Resources: VALLEX and FrameNet

In this section, we briefly characterize two lexical resources used in the project: VALLEX, which takes into account mainly syntactic criteria, and semantically oriented FrameNet.

2.1 VALLEX – Valency Lexicon of Czech verbs

The Valency Lexicon of Czech Verbs, Version 2.5 (VALLEX 2.5)[1] provides information on the valency structure of Czech verbs in their individual senses: primarily, on the number of valency complementations, on their type (labeled by functors), and on their possible morphological forms. VALLEX 2.5 describes 2730 lexeme entries containing about 6460 lexical units (LUs), see [1]. A LU prototypically corresponds to one sense of a verb (only such senses that may be represented by different subcategorization patterns are split into separate LUs). VALLEX 2.5 applies a rather syntactic approach to valency, see [2].

Motivation for Introducing Semantic Classes to VALLEX. Semantic information that reflects the way an individual LU relates to another LU (LUs) plays a key part in NLP tasks, see esp. [3] and [4]. At present, VALLEX does not provide a sufficient insight into semantic relations among LUs.

For illustration, the verbs *prodat*^{pf} 'to sell', ex. (1), and *diktovat*^{impf} 'to dictate', ex. (2), share the same morphosyntactic structure and remain indistinct from each other in VALLEX, in spite of being completely different with respect to their semantics.

(1) *Petr.*ACT *prodal Pavlovi.*ADDR *motorku.*PAT
Eng. *Peter.*ACT *has sold Paul.*ADDR *the motorbike.*PAT
(2) *Ředitel.*ACT *diktoval sekretářce.*ADDR *dopis.*PAT
Eng. *The director.*ACT *has dictated a letter.*PAT *to his secretary.*ADDR

On the one hand, the information on semantic class membership of these verbs allows their differentiation. On the other hand, the semantic classes capturing the relations among LUs enable us to generalize about the morphosyntactic behavior of verbs with similar semantic properties.

[1] `http://ufal.mff.cuni.cz/vallex/2.5/`

2.2 FrameNet

FrameNet[2] is an on-line lexical resource for English. It documents the range of semantic and syntactic combinatory possibilities of each word in each of its senses, see [5]. As to quantitative characteristics, FrameNet contains more than 10,000 lexical units (LUs),[3] pairs consisting of a word and its meaning.

The descriptive framework of FrameNet is based on *frame semantics*. Each LU evokes a particular *semantic frame* (SF) underlying its meaning. Semantic frames consist of semantic arguments, *frame elements*. FrameNet records frame-to-frame relations in the form of a hierarchical network. The relation of 'Inheritance', i.e. the hyperonymy / hyponymy relation, represents the most important one – the semantics of the parent frame corresponds equally or more specifically to the semantics of its child frames.

3 Mapping Semantic Frames from FrameNet to Valency Frames in VALLEX

In this section, we report on our effort to assign the SFs from FrameNet to the VALLEX verbs of communication and verbs of exchange. As the first step, we translated each LU belonging to the selected groups from Czech to English.[4] The total number of translated Czech LUs was 341 for the verbs of communication and 129 for the verbs of exchange (C and E). These LUs correspond to 551 Czech verbs of communication and 325 verbs of exchange, if counting perfective and imperfective counterparts separately.

Two human annotators (A1 and A2) were asked to indicate whether a particular SF evoked by a translated English LU is consistent with a given Czech LU. The annotators were allowed to indicate one SF (labeled as 'Unambiguous assignments of SF') or more than one SF (labeled as 'Ambiguous assignments of SF') for a single Czech LU. The annotators could also conclude that no SF corresponds to a given Czech LU. For the overall statistics, see Table 1.

Table 1. Annotated data size and overall statistics on the annotations of SFs.

	A1 (C / E)	A2 (C / E)
Czech LUs	341 / 129	341 / 129
SFs evoked by English LUs	610 / 171	556 / 166
Unambiguous assignments of SF	143 / 102	165 / 95
Ambiguous assignments of SF	467 / 69	391 / 71
Czech LUs without SFs	83 / 51	83 / 51

[2] `http://framenet.icsi.berkeley.edu/`

[3] For the purposes of this text, we use the same abbreviation LU both for VALLEX and FrameNet because the same concepts are concerned in principle.

[4] The on-line dictionary at `http://www.lingea.cz/` was used.

Inter-Annotator Agreement. Table 2 summarizes the inter-annotator agreement (IAA) and Cohen's κ statistics, see [6], on the total number of used SFs for the verbs of communication and the verbs of exchange. The match of answers related to SFs reaches 85.9% for the verbs of communication and 78.5% for the verbs of exchange. The κ statistics represents an evaluation metric that reflects average pairwise agreement corrected for chance agreement. Both the level 0.82 reached for the verbs of communication and the level 0.73 for the verbs of exchange represent satisfactory results [7].

Table 2. Inter-annotator agreement and κ statistics (considering the annotations of individual SFs for a given Czech LU as independent tasks).

	IAA (C / E)	κ (C / E)	IAA (C + E)	κ (C + E)
Match of SFs	85.9% / 78.5%	0.82 / 0.73	82.2%	0.77

4 Analysis of Semantic Frames Mapping

Statistics on Semantic Frames. 100 SFs in total (SFs regardless of the inter-annotator agreement) were mapped to 341 verbs of communication. The most frequently assigned SFs to the verbs of communication include: 'Statement', 'Request', 'Telling', 'Communication_manner', 'Reporting', etc.

52 SFs in total (regardless of the inter-annotator agreement) were assigned to 129 verbs of exchange. The following SFs belong to the most frequently assigned ones: 'Giving', 'Getting', 'Exchange', 'Commerce_pay', 'Theft', etc.

The number of SFs is significantly lower, if taking into account only those in which the annotators concurred – 69 for the verbs of communication and 31 for the verbs of exchange.

Ambiguous Assignment of Semantic Frames. Ambiguous annotations draw attention to the divergence in granularity of word sense disambiguation adopted by VALLEX and FrameNet. The different level of granularity represents a great setback in making one-to-one correspondences between LUs from VALLEX and those from FrameNet. In Section 5, we propose a method to overcome this difficulty.

Let us focus on the cases in which two (or more) SFs mapped to a single Czech LU are connected by the hierarchical relation of 'Inheritance' – these cases reveal the finer granularity of senses applied in FrameNet. For instance, the SFs 'Getting' and 'Earnings_and_losses' are assigned to the single Czech LU *mít* 'to get / to earn' as in *Who did you get it from?* and *He earns five thousand per month*, respectively. The SF 'Earnings_and_losses' is a descendant of the SF 'Getting' in the relation 'Inheritance': Although the LU 'to earn' from the SF 'Earnings_and_losses' is semantically more specified, it inherits semantic properties from the LU 'to get' evoking the SF 'Getting'.

FrameNet data can also be used for checking word sense disambiguation in VALLEX. The ambiguous annotations of SFs that do not arise from the finer granularity may reveal mistakes in word sense disambiguation. For instance, the SFs 'Grant_permission' and 'Permitting' are assigned to the Czech LU *dovolit*^{pf}, *dovolovat*^{impf} 'to allow', as in *Peter has allowed me to smoke here* and *This program allows data checking*, respectively. Although the LUs appear to be semantically close, the SFs evoked by them are not in the relation of 'Inheritance'. Thus this Czech LU represents a candidate for being split into two distinct senses.

5 Exploiting FrameNet for Enhancing VALLEX with Semantic Classes

The hierarchical network of relations between SFs in FrameNet plays a key role in the classification of Czech LUs. The relation of 'Inheritance' is of major importance as each child frame inherits semantic properties from its parent frame(s).

We made use of the upper levels of the relation of 'Inheritance' for grouping LUs into coherent semantic classes: we mapped the ancestor frames from appropriate levels of the relation of 'Inheritance' to Czech LUs. This method allows us to surmount the problem with the coarser granularity of verb senses in VALLEX, see Section 4.

However, the top levels of the relation of 'Inheritance' cannot be exploited as they are occupied by abstract and non-lexical SFs, as e.g. 'Intentionally_act' and 'Reciprocality', respectively. Similarly, the top SFs describing only a very general event, as the SF 'Event' (which may be understood as the core of all events), are excluded.

For instance, the SF 'Giving' represents the proper level in the 'Inheritance' hierarchy. Prototypically, Czech LUs to which descendant SFs of the SF 'Giving' are assigned (see Figure 1) are included in the semantic class 'Giving'.

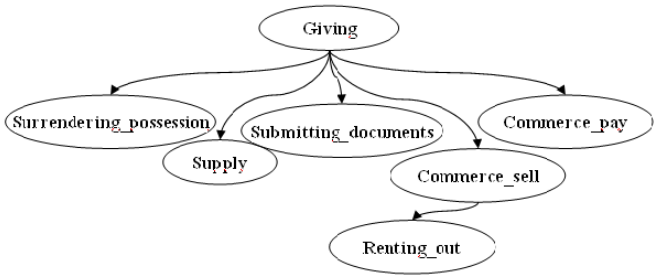

Fig. 1. The upper levels of the relation 'Inheritance' of the SF 'Giving'.

However, if Czech LUs exhibit different morphosyntactic properties than the LUs to which the ancestor SF is assigned, we map the SF from the lower level of the relation of 'Inheritance'. For instance, *odměnit*^{pf} 'to reward', to which the SF 'Supply' is mapped, differs in its morphosyntactic properties from the LUs to which the SF 'Giving' is assigned (as e.g. *odevzdat*^{pf}, 'to surrender', *prodat*^{pf} 'to sell', etc.). Thus 'Supply' is considered as a candidate for another semantic class.

Semantic Classes for Verbs of Exchange. 31 SFs[5] assigned to the verbs of exchange correspond to 16 SFs on the upper levels of the 'Inheritance' hierarchy. However, half of them cannot be exploited: six of these SFs have not yet been linked by the 'Inheritance' relation and two SFs do not appear to be relevant.[6] The remaining eight SFs were accepted as semantic classes (1.-8.). We made use of two other SFs 'Taking' and 'Supply' from the lower levels of the 'Inheritance' hierarchy, due to syntactic differences of the LUs to which these SFs are assigned (9.-10.). To conclude, we obtained the following candidates for semantic classes:

1. **'Giving'**: *odevzdat*^{pf}, 'to surrender', *prodat*^{pf} 'to sell', etc.,
2. **'Getting'**: *koupit*^{pf} 'to buy', *přijímat*^{impf} 'to receive', etc.,
3. **'Replacing'**: *nahradit*^{pf} 'to replace', etc.,
4. **'Exchange'**: *měnit*^{impf} 'to exchange', *vyměnit*^{pf} 'to exchange', etc.,
5. **'Robbery'**: *okrást*^{pf} 'to rob', *připravit*^{pf} 'to rob', etc.,
6. **'Hiring'**: *najmout*^{pf} 'to hire', etc.,
7. **'Transfer'**: *postoupit*^{pf}, 'to hand over', *připisovat*^{impf} 'to transfer', etc.,
8. **'Frugality'**: *utratit*^{pf} 'to waste', etc.,
9. **'Taking'**: *krást*^{impf} 'to steal', *vzít*^{pf} 'to take', etc.,
10. **'Supply'**: *odměnit*^{pf} 'to reward', *opatřovat*^{impf} 'to provide', etc.

Semantic Classes for Verbs of Communication. The SFs corresponding to the verbs of communication are finer-grained than those corresponding to the verbs of exchange. Moreover, only 23 SFs from 69 SFs[7] assigned to the verbs of communication are connected by the relation of 'Inheritance' in FrameNet at present.

The SFs from the hierarchy in Figure 2 belong to the most often assigned – they cover almost 37% of verbs of communication. However, we did not use the top level SF 'Communication' for all the LUs because it is too general.

We mapped the SF 'Communication' as the semantic class to the LUs to which its descendant SFs 'Communication_noise', 'Communication_manner', 'Gesture', and 'Summarizing' were assigned. With respect to morphosyntactic properties of Czech verbs of communication, we accepted also two SFs from the lower

[5] We took into account only the cases in which the annotators concurred.

[6] The SF 'Rewards_and_punishment' is mapped only to the LU *odměnit*^{pf} 'to reward'. However, as this LU shares the morphosyntactic and semantic properties with the LUs to which the SF 'Supply' is assigned, we include it in this semantic class. The SF 'Agree_or_refuse_to_act' was assigned only to the LU *odepřít*^{pf}. However, we leave this SF aside because it does not express exchange.

[7] As in the case of the verbs of exchange, we took into account only the cases in which the annotators concurred.

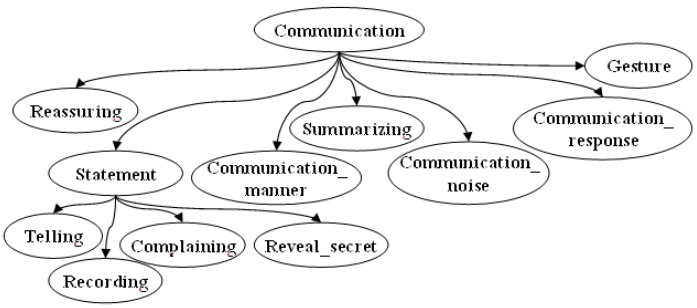

Fig. 2. The upper levels of the relation 'Inheritance' of the SF 'Communication'.

level of this hierarchy – 'Statement' and 'Communication_response'.[8] As a result, we obtained three semantic classes based on the above given hierarchy:

1. **'Communication'**: *šeptat*^{impf} 'to whisper', *zpívat*^{impf} 'to sing', etc.,
2. **'Statement'**: *oznámit*^{pf}, 'to announce', *tvrdit*^{impf} 'to claim', etc.,
3. **'Communication_response'**: *odvětit*^{pf} 'to reply', etc.

Three other semantic classes arose from the upper levels of the 'Inheritance' hierarchy: (4) 'Judgment_communication', covering also the LUs to which the SFs 'Judgment_direct_address' and 'Bragging' are assigned , (5) 'Chatting', which includes the LUs to which the SFs 'Discussion' and 'Quarreling' are assigned, and (6) 'Prohibiting', which involves the LUs to which the SF 'Permitting' is assigned as well.

4. **'Judgment_communication'**: *vyčítat*^{impf} 'to reproach', *děkovat*^{impf} 'to thank', *obvinit*^{pf} 'to accuse', etc.,
5. **'Chatting'**: *diskutovat*^{biasp}, 'to discuss', *hádat se*^{impf} 'to quarrel', etc.,
6. **'Prohibiting'**: *zakázat*^{pf}, 'to prohibit', *povolit*^{pf} 'to allow', etc.

The remaining SFs frequently assigned to the verbs of communication, such as 'Request'(7),[9] 'Reporting' (8), and 'Commitment' (9), were accepted as further candidates for separate semantic classes, despite not being linked by the 'Inheritance' relation. Apparently, Czech LUs to which these SFs were assigned exhibit distinct syntactic behavior – the small differences in morphemic forms are left out of account.

7. **'Request'**: *nabádat*^{impf} 'to urge', *poručit*^{pf} 'to order', *prosit*^{impf} 'to ask', *přemlouvat*^{impf} 'to urge', *vyzvat*^{pf} 'to ask', etc.,

[8] The SF 'Reassuring' was never assigned.

[9] The SF 'Request' was frequently mapped to the LUs along with the SF 'Attempt_suasion'. The LUs to which these SFs were assigned are similar in many aspects. Thus we take into account only one semantic class 'Request'.

8. **'Reporting'**: *hlásit*^{impf}, 'to report', *říci*^{pf} 'to tell', *udat*^{pf} 'to report', etc.,
9. **'Commitment'**: *hrozit*^{impf} 'to threaten', *slíbit*^{impf}, 'to promise', *přísahat*^{impf} 'to vow', etc.

As a result, more than 68% of verbs of communication and almost 98% of verbs of exchange in VALLEX are grouped into the semantic classes listed above. In the future, the verbs with no assigned SFs (due to the translated English LU not being covered in FrameNet) will be further examined in order to be included into the appropriate semantic class. Furthermore, we intend to add more classes to the list in agreement with the progress made in FrameNet.

Conclusion

We have presented an experiment in enriching a valency lexicon of Czech verbs, VALLEX, with semantic classes. We have mapped the FrameNet semantic frames to Czech verbs of communication and exchange. We have attained a satisfactory inter-annotator agreement, which proves the feasibility of this task.

We have exploited semantic frames from the upper levels of the relation of 'Inheritance' for classifying the verbs into semantic classes. As a result, we have established nine classes for the verbs of communication, such as 'Statement', 'Request', 'Reporting', etc., and ten classes for the verbs of exchange, such as 'Giving', 'Getting', 'Exchange', etc. These classes cover more than 68% of verbs of communication and almost 98% of verbs of exchange.

In the future, we plan to experiment with semantic frames for other groups of verbs, especially for the verbs of motion, transport, mental action, psych verbs, etc. Moreover, we intend to make use of FrameNet frame elements as semantic labels for verb arguments.

References

1. Žabokrtský, Z., Lopatková, M.: Valency Information in VALLEX 2.0: Logical Structure of the Lexicon. The Prague Bulletin of Mathematical Linguistics 87 (2007) 41–60
2. Panevová, J.: Valency Frames and the Meaning of the Sentence. In Luelsdorff, P.L., ed.: The Prague School of Structural and Functional Linguistics. John Benjamins, Amsterdam-Philadelphia (1994) 223–243
3. Loper, E., Yi, S., Palmer, M.: Combining Lexical Resources: Mapping between PropBank and VerbNet. In: Proceedings of the 7th International Workshop on Computational Linguistics. (2007)
4. Kingsbury, P., Palmer, M., Marcus, M.: Adding Semantic Annotation to the Penn TreeBank. In: Proceedings of the Human Language Technology Conference. (2002)
5. Ruppenhofer, J., Ellsworth, M., Petruck, M.R.L., Johnson, C.R., Scheffczyk, J.: FrameNet II: Extended Theory and Practice. (2006) http://framenet.icsi.berkeley.edu/book/book.html.
6. Carletta, J.: Assessing agreement on classification tasks: The Kappa statistic. Computational Linguistics **22** (1996) 249–254
7. Krippendorf, K.: Content Analysis: An Introduction to its Methodology. Sage Publications (1980)