Valency Lexicon of Czech Verbs VALLEX: Recent Experiments with Frame Disambiguation*

Markéta Lopatková, Ondřej Bojar, Jiří Semecký,
Václava Benešová, and Zdeněk Žabokrtský

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics, Charles University, Prague
{lopatkova,bojar,semecky,benesova,zabokrtsky}@ufal.mff.cuni.cz

Abstract. VALLEX is a linguistically annotated lexicon aiming at a description of syntactic information which is supposed to be useful for NLP. The lexicon contains roughly 2500 manually annotated Czech verbs with over 6000 valency frames (summer 2005). In this paper we introduce VALLEX and describe an experiment where VALLEX frames were assigned to 10,000 corpus instances of 100 Czech verbs – the pairwise inter-annotator agreement reaches 75%. The part of the data where three human annotators agreed were used for an automatic word sense disambiguation task, in which we achieved the precision of 78.5%.

1 Introduction

A verb is traditionally considered to be the center of the sentence, and description of syntactic and syntactic-semantic behavior of verbs is a substantial task for linguists. Theoretical aspects of valency are challenging. Moreover, valency information stored in a lexicon (as valency properties are multifarious and cannot be described by general rules) belongs to the core information for any rule-based task of NLP (from lemmatization and morphological analysis through syntactic analysis to such complex tasks as e.g. machine translation).

There are tens of different theoretical approaches, tens of language resources and hundreds of publications related to the study of verbal valency in various natural languages. It goes far beyond the scope of this paper to give an exhaustive survey of all these enterprises – [1] gives a survey and a short characteristics of the most prominent projects.

The present paper is structured as follows: in Section 2 we summarize the basic properties of the lexicon VALLEX, in Section 3 we describe the human-annotated data where corpus occurrences of selected verbs are assigned to valency frames, in Section 4 we report the experiment with automatic frame assignment.

2 Valency Lexicon of Czech Verbs VALLEX

The VALency LEXicon of Czech verbs (VALLEX in the sequel) is a collection of linguistically annotated data and documentation, resulting from an attempt at formal

* The research reported in this paper has been partially supported by the grant of Grant Agency of Czech Republic No. 405/04/0243 and by the projects of Information Society No 1ET100300517 and 1ET101470416.

V. Matoušek et al. (Eds.): TSD 2005, LNAI 3658, pp. 99–106, 2005.

description of valency frames of Czech verbs. VALLEX version 1.0 was publicly released in autumn 2003[1]. VALLEX 1.0 contained roughly 1400 verbs with 4000 valency frames. At this moment, the latest version of VALLEX data contains roughly 2500 verbs with more than 6000 valency frames. All verb entries are created manually. Manual annotation and accent put on consistency of annotation are markedly time consuming and limit the speed of quantitative growth, but guarantees a significant rise of quality.

VALLEX is closely related to Prague Dependency Treebank (PDT)[2]. Both PDT and VALLEX are based on Functional Generative Description of Czech (FGD), being developed by Petr Sgall and his collaborators since the 1960s (see [3], valency theory within FGD esp. in [4]). Applying the principles of FGD to a huge amount of data means a great opportunity to verify and expand the theory, to refine the functional criteria set up. The modification of 'classical' FGD valency theory is used as the theoretical background in VALLEX 1.0 (see [5] for a detailed description of the framework).

On the topmost level, VALLEX[3] consists of **word entries** corresponding to complex units, verb lexemes (the VALLEX entries for the verbs *odpovídat* and *odpovídat se* is shown in Figure 1). The particular word entry is characterized by the **headword lemma**, i.e. the infinitive form of the respective verb (including the reflexive particle if it exists) and its **aspect** (perfective, imperfective or biaspectual). The tentative term **base lemma** denotes the infinitive of the verb, excluding the reflexive particle (i.e. the output of a morphological analysis).

Each word entry is composed of a non-empty sequence of **frame entries** relevant for the headword lemma. The frame entries (marked with subscripts in VALLEX) roughly correspond to individual senses of the headword lemma. The particular word entry is characterized by a **gloss** (i.e. verb or paraphrase roughly synonymous with the given frame/sense) and by **example(s)** (i.e. sentence fragment(s) containing the given verb used with the given valency frame). The core valency information is encoded in the **valency frame**.

Each valency frame consists of a set of **valency members / frame slots**, each corresponding to an (either required or specifically permitted) complementation of the given verb. The information on a particular valency member includes the following points:

– **'Functor'** expresses the type of relation between the verb and its complementation[4]. Complementations are divided into (i) inner participants / arguments (like Actor, Patient and Addressee for the verb *přinést*$_1$ [to bring], as in *někdo*.ACT *přinese něco*.PAT *někomu*.ADDR [sbd brings st to sbd] or Actor, Patient and Effect for the verb *jmenovat*$_3$ [to nominate], as in *někdo*.ACT *jmenuje někoho*.PAT *něčím*.EFF

[1] `http://ckl.mff.cuni.cz/zabokrtsky/vallex/1.0/`

[2] However, VALLEX is not to be confused with a larger valency lexicon PDT-VALLEX created during the annotation of PDT, see [2]. PDT-VALLEX contains more verbs (5500 verbs), but only frames occurring in PDT (over 9000 frames), whereas in the more complex VALLEX the verbs are analyzed in all their meanings. In addition, richer information is assigned to particular valency frames.

[3] Detailed description can be found in [6].

[4] The complete list of functors used in VALLEX together with English examples can be found in [6].

odpovídat (imperfective)

1 odpovídat$_1$ ∼ odvětit [answer; respond]
- frame: ACT_1^{obl} ADDR_3^{obl} $\mathrm{PAT}_{na+4,4}^{opt}$ $\mathrm{EFF}_{4,aby,ať,zda,že}^{obl}$ MANN^{typ}
- example: *odpovídal mu na jeho dotaz pravdu / že …* [he responded to his question truthfully / that …]
- asp.counterpart: odpovědět$_1$ pf.
- class: communication

2 odpovídat$_2$ ∼ reagovat [react]
- frame: ACT_1^{obl} $\mathrm{PAT}_{na+4}^{obl}$ MEANS_7^{typ}
- example: *pokožka odpovídala na včelí bodnutí zarudnutím* [the skin reacted to a bee sting by turning red]
- asp.counterpart: odpovědět$_2$ pf.

3 odpovídat$_3$ ∼ mít odpovědnost [be responsible]
- frame: $\mathrm{ACT}_1^{obl}\mathrm{ADDR}_3^{obl}\mathrm{PAT}_{za+4}^{opt}\mathrm{MEANS}_7^{typ}$
- example: *odpovídá za své děti; odpovídá za ztrátu svým majetkem* [she is responsible for her kids]

4 odpovídat$_4$ ∼ být ve shodě [match]
- frame: $\mathrm{ACT}_{1,že}^{obl}\mathrm{PAT}_3^{obl}\mathrm{REG}_7^{typ}$
- example: *řešení odpovídá svými vlastnostmi požadavkům* [the solution matches the requirements]

odpovídat se (imperfective)

1 odpovídat se$_1$ ∼ být zodpovědný [be responsible]
- frame: $\mathrm{ACT}_1^{obl}\mathrm{ADDR}_3^{obl}\mathrm{PAT}_{z+2}^{obl}$
- example: *odpovídá se ze ztrát* [he answers for the losses]

Fig. 1. VALLEX entries for the base lemma *odpovídat (answer, match)*.

[sbd nominates sbd as sbd]) and (ii) free modifications (adjuncts) as Time, Location, Manner and Cause[5].

- Possible **morphemic form(s)** – each complementation can be expressed by a limited set of morphemic means (pure or prepositional cases, subordinated clauses or infinitive constructions are the most important); possible morphemic form(s) are specified either explicitly (as a list of forms attached to a particular slot) or implicitly[6].
- **'Type'** – the following types of complementations are distinguished: obligatory (in the deep (tectogrammatical) structure) and optional for inner participants ('obl' and 'opt'), and obligatory and typical ('typ') for free modifications.

In addition to this obligatory information, also optional attributes may appear in each frame: flag for idiom, list of aspectual counterpart(s), information on control, affiliation to a syntactic-semantic class:

[5] Here we are leaving aside a small group of complementations on the border-line between inner participants and free modifications, quasi-valency complementations, see [5].

[6] The set of possible forms is implied by the functor of the complementation, see [6].

- **Flag for idiom** – VALLEX describes primary or usual meanings of verbs, however some very frequent idiomatic frames[7] are included as well. They are marked by idiomatic flag and include lemmas of words in the phraseme.
- **Aspectual counterpart** – aspectual counterpart(s) need not be the same for all senses of the given verb; if they exist, they are listed in particular frame entries[8] (see figure 1).
- **Control** – if a verb has a complementation in an infinitive form (regardless its functor), the valency member of the head verb that would be the subject of this infinitive is marked.
- **Syntactic-semantic classes** – particular frame entries are tentatively sorted into classes. Constructed in a 'bottom-up way', these classes are based on deep analysis of mainly syntactic properties of verbs in their particular senses. For the time being, 24 big groups involving next to half of the verb frames have been established[9].

3 VALEVAL

VALEVAL[10] is a lexical sampling experiment with VALLEX 1.0 for which 109 base lemmas from VALLEX 1.0 were selected. For each lemma 100 random sample sentences were extracted from CNC. See [7] for more details and examples.

Three human annotators in parallel were asked to choose the most appropriate verb entry and the frame for the extracted sentence within a context of the three preceding sentences. The annotators had also an option to indicate that the particular sentence is not a valid example (e.g. due to a tagging error) of the annotated lemma at all or that they got completely confused by the given context. A valid answer indicates a verb entry and a frame entry index. Optionally, a remark that the corresponding frame was missing could have been given instead of the frame entry index. If the annotators were not able to decide on a single answer, they have been given the possibility of assigning more than one valid answer (labelled as 'Ambiguous annotations' in Table 1). Also, a special flag could be assigned to a valid answer to indicate that the annotator is not quite sure (labelled as 'Uncertain annotations').

3.1 Inter-annotator Agreement

Table 2 summarizes inter-annotator agreement (IAA) and Cohen's κ statistic [9] on the 10256 annotated sentences. The symbol Ø indicates plain average calculated over base lemmas, wØ stands for average weighted by frequency observed in CNC. Considering all the three parallel annotations, the exact match of answers reaches 61% (weighted)

[7] Idiomatic frame is tentatively characterized either by a substantial shift in meaning (with respect to the primary sense), or by a small and strictly limited set of possible lexical values in one of its complementations.

[8] Iterative verbs occur in entries of the corresponding non-iterative verbs, but they have no own word entries.

[9] However rough these classes are, they serve for controlling the consistency of annotation.

[10] Inspired by SENSEVAL ([8]), a word sense disambiguation task, VALEVAL aims at valency frame disambiguation.

Table 1. Annotated data size and overall statistics about the annotations.

Lemmas annotated	109
Sentences annotated	10256
Parallel annotators	3
Total annotations	30765 (100%)
Uncertain annotations	1045 (3.4%)
Ambiguous annotations	703 (2.3%)
Marked as invalid example	172 (0.6%)
Annotator got confused	90 (0.3%)
Marked as missing frame	1673 (5.4%)

or 67% (unweighted). If the 'uncertainty' flags are disregarded, we find out that the agreement rises to 66% or 70%, respectively. In other words, annotators agree on the most plausible answer, even if they are not quite sure. If only such sentences where none of the annotators doubted are taken into account, the exact match reaches 68% or 74% (this comprises 90.5% of the sentences).

The κ statistic compensates IAA for agreement by chance. The level of 0.5 to 0.6 we achieve is generally considered as a *moderate agreement*, while 0.6 to 0.8 represents *significant agreement*. This moderate agreement is not an unsatisfactory result compared to other results such as [10], who reports pairwise IAA for French verbs between 60% and 65% and κ of 0.41.

Table 2. Inter-annotator agreement and κ.

	Match of 3 Annotators				**Average Pairwise Match**			
	IAA [%]		κ		IAA [%]		κ	
	wØ	Ø	wØ	Ø	wØ	Ø	wØ	Ø
Exact	61.4	66.8	0.52	0.54	70.8	74.8	0.54	0.54
Ignoring Uncertainty	65.9	69.8	0.58	0.59	74.8	77.7	0.60	0.59
Where All Were Sure	68.2	73.7	0.58	0.62	76.7	80.9	0.61	0.64

Average pairwise IAA is provided to allow for a rough comparison with some cited results, although the specific circumstances are not always directly comparable. [11] achieve an IAA for Czech verbs of 45% to 64%. For Japanese verbs, IAA of 86.3% is achieved by [12]. [13] report IAA of 71% for Senseval-2 English verbs tagged with WordNet synsets. Grouping some senses together to form a more coarse grained sense inventory allowed the authors to improve the IAA to 82%.

4 Automatic Frame Disambiguation

4.1 Data Source: 'Golden VALEVAL'

VALLEX frames correspond to verb senses (meanings). From this perspective, performing word sense disambiguation (WSD) of Czech verbs means choosing the most

Table 3. Baselines for WSD on 8066 'Golden VALEVAL' sentences for 108 lemmas.

	wØ	Ø
Entropy	1.54	1.28
VALLEX frames per lemma	12.46	7.61
Seen frames per lemma	5.85	4.85
10-fold Baseline WSD Accuracy	59.79%	66.19%

appropriate frame. 'Golden VALEVAL' is a corpus suitable for evaluating frame disambiguation. It comprises 8066 VALEVAL sentences covering 108 base lemmas where there was exact agreement across the annotators or a single answer was selected in a postprocess annotation aimed at eliminating clear typing errors and misinterpretations.

The difficulty of the WSD task is apparent from Table 3 looking at the (weighted or unweighted average) number of available frames per base lemma and entropy. The number of frames per lemma is estimated both from the whole VALLEX ('VALLEX frames per lemma') as well as from the set of actually observed frames in the golden VALEVAL corpus ('Seen frames per lemma').

The baseline accuracy is achieved by choosing the most frequent frame for a given lemma. The baseline was estimated by a 10-fold cross-validation (the most frequent frame is learned from 9/10 of the data and the unseen 1/10 is used to estimate the accuracy, the average result from 10 runs of the estimation is reported).

For purposes of further experiments, Golden VALEVAL was automatically tagged, lemmatized and enriched with surface syntactic structures automatically assigned by the Czech version of the parser reported in [14]. After the exclusion of unparsed sentences, 6666 sentences remained for our task.

4.2 Method and Selected Features

For an automatic selection of the VALLEX frame to which a given verb occurrence belongs, we generated a vector of features for each occurrence. We evaluated the decision tree machine learning method available in C5 toolkit[11]. 10-fold cross-validation was used for evaluation.

We experimented with several features containing information about the context of the verb. The following list describes different groups of features:

- Morphological: purely morphological information about lemmas in a 5-word window centered around the verb. Czech positional morphological tags (used also in PDT) contain 15 categories and all of these were taken as individual features, counting 75 features altogether.
- Syntax-based: information gained from the dependency tree of the sentence, including mostly Boolean information about morphological and lexical characteristics of dependent words (e.g. presence of a noun or a nominative pronoun in a given case dependent on the verb, presence of a given preposition with a given case dependent on the verb).

[11] `http://www.rulequest.com/see5-info.html`

4.3 Results

Weighting the accuracy by the number of sentences in our training set (labelled as Ø in Table 4), we gained 73.9% accuracy for morphological features and 78.5% accuracy for syntax-based features, respectively, compared to baseline 67.9% (baseline for the 6666 parsed sentences). Weighting the accuracy by the lemma frequency observed in the Czech National Corpus (labelled as wØ), the accuracy dropped to 67.1% for the morphological features and 70.8% for syntax-based features respectively, compared to baseline 63.3%.

Table 4. Accuracy of frame disambiguation.

	wØ	Ø
Baseline	63.3%	67.9%
Morphological	67.1%	73.9%
Syntax-based	70.8%	78.5%

The syntax-based features alone led to better results, and even the combination of both of the types of features did not bring any improvement. This could happen because the morphological information is already included in the syntax-based features (as they contain information mainly about morphological characteristics of syntactically related words) and because the syntactic structure of the sentence depicts enough information to achieve the rate of disambiguation which can be obtained using this method.

5 Conclusions and Future Work

We have presented the current state of building valency lexicon of Czech verbs VALLEX. We have also described the VALEVAL experiment which allowed us to improve consistency of selected VALLEX entries and provided us with golden standard data for WSD task. The first results in WSD are reported.

In future we plan to extend VALLEX in both qualitative aspects (e.g. description of alternations and types of reflexivity) and quantitative aspects. We will continue the WSD experiments, we intend to incorporate features based on WordNet classes and animacy.

References

1. Žabokrtský, Z.: Valency Lexicon of Czech Verbs. PhD thesis, Faculty of Mathematics and Physics, Charles University in Prague (2005) in prep.
2. Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V., Pajas, P.: PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In: Proceedings of The Second Workshop on Treebanks and Linguistic Theories. Volume 9 of Mathematical Modeling in Physics, Engineering and Cognitive Sciences., Vaxjo University Press (2003) 57–68

3. Sgall, P., Hajičová, E., Panevová, J.: The Meaning of the Sentence in Its Semantic and Pragmatic Aspects. D. Reidel Publishing Company, Dordrecht (1986)
4. Panevová, J.: Valency Frames and the Meaning of the Sentence. In Luelsdorff, P.L., ed.: The Prague School of Structural and Functional Linguistics, Amsterdam-Philadelphia, John Benjamins (1994) 223–243
5. Lopatková, M.: Valency in the Prague Dependency Treebank: Building the Valency Lexicon. Prague Bulletin of Mathematical Linguistics **79–80** (2003) 37–60
6. Žabokrtský, Z., Lopatková, M.: Valency Frames of Czech Verbs in VALLEX 1.0. In: Frontiers in Corpus Annotation. Proceedings of the Workshop of the HLT/NAACL Conference. (2004) 70–77
7. Bojar, O., Semecký, J., Benešová, V.: VALEVAL: Testing VALLEX Consistency and Experimenting with Word-Frame Disambiguation. Prague Bulletin of Mathematical Linguistics **83** (2005)
8. Edmonds, P.: Introduction to Senseval. ELRA Newsletter **7** (2002)
9. Carletta, J.: Assessing agreement on classification task: The kappa statistics. Computational Linguistics **22** (1996) 249–254
10. Véronis, J.: A study of polysemy judgements and inter-annotator agreement. In: Programme and advanced papers of the Senseval workshop, Herstmonceux Castle (England) (1998) 2–4
11. Hajič, J., Holub, M., Hučínová, M., Pavlík, M., Pecina, P., Straňák, P., Šidák, P.: Validating and Improving the Czech WordNet via Lexico-Semantic Annotation of the Prague Dependency Treebank. In: Proceedings of LREC 2004. (2004)
12. Shirai, K.: Construction of a Word Sense Tagged Corpus for SENSEVAL-2 Japanese Dictionary Task. In: Proceedings of LREC 2002. (2002) 605–608
13. Babko-Malaya, O., Palmer, M., Xue, N., Joshi, A., Kulick, S.: Proposition Bank II: Delving Deeper. In: Frontiers in Corpus Annotation. Proceedings of the Workshop of the HLT/NAACL Conference. (2004) 17–23
14. Charniak, E.: A Maximum-Entropy-Inspired Parser. In: Proceedings of NAACL-2000, Seattle, Washington, USA (2000) 132–139