VALEVAL: Testing VALLEX Consistency and Experimenting with Word-Frame Disambiguation

Ondřej Bojar, Jiří Semecký and Václava Benešová

Abstract

VALLEX is a valency lexicon of Czech verbs. We briefly introduce VALLEX and then we describe and evaluate the VALEVAL experiment: annotation of 10256 corpus instances of 109 Czech verbs with valency frames. The inter-annotator agreement of three parallel annotations ranges from 61% to 74% and κ from 0.52 to 0.62. More than 8000 sentences are now available as the "golden VALEVAL" for word-sense disambiguation experiments. In out first attempts using morphological and syntactic information, we achieve the accuracy of 70% to 80%.

1 Introduction

Verbs are central to the structure of a sentence. Capturing syntactico-semantic properties such as valency frames of individual verbs is important for many NLP tasks, such as lemmatisation, tagging, syntactic analysis (parsing), semantic analysis, information retrieval, as well as for more complex NLP goals including machine translation, question answering or automatic inference.[1]

It is well known that verb valency cannot be described adequately by means of general rules and that a lexicalist approach si required. There are dozens of different theoretical approaches, tens of language resources and hundreds of publications related to the study of verb valency in various natural languages. Unfortunately, only the most prominent projects like **?**) or **?**) are usually cited and to the best of our knowledge there is no sufficiently representative comparison available in the literature so far. It goes far beyond the scope of this paper to give an exhaustive survey of such enterprises; e.g., see **?**) for a review.

This paper is organized as follows: Section 2 gives a brief introduction to the VALLEX and comments on the difficulty of frame assignment using VALLEX and also the frequencies of verbs covered in VALLEX. Section 3 describes our experiment of annotating sample verb occurrences with VALLEX entries. Section 4 is devoted to the description of properties of golden standard data that were created by VALEVAL annotation. We employ the data in a frame disambiguation experiment which is described in section 5. VALEVAL provided us with a lot of feedback on the quality of VALLEX entries, a brief summary of the problems spotted is given in Section 6.

2 VALLEX

2.1 Introducing VALLEX

VALLEX[2] (**?**) or **?**)) is a valency lexicon for Czech being built manually since 2001. VALLEX aims at capturing syntactico-semantic properties of Czech verbs according to the valency theory by **?**) and within the framework of the Functional Generative Description (FGD, **?**)). **?**) describe the valency theory in detail and incorporate minor changes in order to reflect properly all observations made by VALLEX developers.

[1] Consult **?**) for several examples.

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

VALLEX focuses on individual lexical items and aims at capturing the whole complexity once an item is covered, i.e. VALLEX entries are intended to be complete with all their senses listed (possibly with the exception of idiomatic expressions).[3] The version 1.0 of VALLEX is based on 1000 most frequent Czech verbs; after aspectual and reflexive counterparts of the verbs have been added (VALLEX is closed under the relation "aspectual pair"), a total of 1400 Czech verbs are covered by VALLEX-1.0.

2.2 The Structure of VALLEX

At the topmost level, VALLEX is a list of VERB ENTRIES[4], see Figure 1 for an example of two of them. The verb is characterized by its HEADWORD LEMMA (including a reflexive particle *se* or *si*, if present) and its aspect or several spelling variants of the headword lemma. Every verb entry includes one or more VALENCY FRAMES of the verb roughly corresponding to its senses. Every valency frame consists of a set of VALENCY SLOTS characterizing complementations of the verb. Each slot describes both the type of the syntactico-semantic relation between the verb and its complementation (by means of a "tectogrammatical functor", such as Actor, Patient, Direction; see FGD) and all allowed surface realizations ("morphemic forms") of the verb complementation (e.g. the required preposition and case or the subordinating conjunction for subclauses). The slot also indicates obligatoriness of the complementation. Each frame is equipped with a short gloss and an example in order to help human annotators to distinguish among the frames. Aspectual counterparts of the verb are not assigned to the verb entry as a whole but to the individual frames: a frame of a verb contains a link to a frame of its aspectual counterpart, if appropriate.

The operational criteria on when to create a new frame entry of a verb are described in **?**). Roughly speaking, a frame entry corresponds to a "sense" of the verb based primarily on (deep) syntactic observations.

The tentative term BASE LEMMA denotes the infinitive of the verb, excluding a possible reflexive particle and homograph distinction, e.g. *odpovídat* is the base lemma for the verbs *odpovídat* and *odpovídat se*. The base lemma is assumed to be a part of an output of a morphological analysis of text.

2.3 Difficulty of Frame Assignment

As indicated in Table 1, there is a big difference between assigning a frame to a known verb entry and assigning a frame to a base lemma if the verb entry is still unknown. The average number of frames per verb entry in VALLEX is 2.7 and it varies from a single frame to at most 26 frames[5] assigned to the verb *brát* (take, pick up). Taking only the base lemma as an input, the average number of frames that could be assigned rises to 3.9.

	Verb Entries	Base Lemmas
Avg. frames	2.7	3.9
Max. frames	26	33
Total #	1437	1080
# with single frame	536	207

Table 1: Verb entries and lemmas in VALLEX.

The main reason for such a growth in frame ambiguity is the reflexivity of verbs. VALLEX treats reflexive and non-reflexive counterparts (e.g. *bavit* (amuse) vs. *bavit se* (have fun, talk)) as separate

[3] VALLEX should not be confused with PDT-VALLEX (**?**), a lexicon covering only the frames observed in Prague Dependency Treebank.

[4] Due to the lack of space we can only briefly summarise the key terms. Please consult **?**) for a detailed description, examples and explanation of all the terms not defined here.

[5] Cf. Section 6 to learn more about incompleteness of VALLEX with respect to idioms.

odpovídat (imperfective)

1 odpovídat$_1$ ∼ odvětit (answer; respond)
- frame: ACT_1^{obl} ADDR_3^{obl} $\text{PAT}_{na+4,4}^{opt}$ $\text{EFF}_{4,aby,ať,zda,že}^{obl}$ 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} 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: $\text{ACT}_1^{obl}\text{ADDR}_3^{obl}\text{PAT}_{za+4}^{opt}\text{MEANS}_7^{typ}$
- example: *odpovídá za své děti; odpovídá za ztrátu svým majetkem (she is responsible for her kids)*

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

odpovídat se (imperfective)

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

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

verb entries. However, this distinction cannot be made from the base lemma itself. Table 2 gives a detailed overview: there are 627 base lemmas in VALLEX that never take any reflexive particle. However, there are 326 base lemmas that create a verb entry when accompanied by reflexive particle and another entry when not accompanied by it. Verbs that require the reflexive particle (either *se* or *si* or both) to accompany their base lemma are not very common.

Reflexive variants	# Base lemmas	Example
no particle	627	důvěřovat (trust)
possible *se*	326	držet (hold)
obligatory *se*	53	dívat (look)
possible *si*	40	hrát (play)
possible *se* or *si*	29	dát (give)
obligatory *si*	6	stěžovat (complain)

Table 2: Reflexive variants of base lemmas in VALLEX

The question whether a *se* or *si* found near a verb in a sentence denotes the verb's reflexive particle is quite complex and cannot be performed automatically without a syntactically annotated corpus. Using Prague Dependency Treebank (PDT, **?**)) we observe that given a base lemma with at least 25 occurrences in PDT, the tendency is to prefer either one of its reflexive variants or the non-reflexive variant. Figure 2 illustrates our observation both on PDT as well as on our VALEVAL data (see Section 3 below). By RELATIVE REFLEXIVITY of a base lemma we mean the percentage of observations where the base lemma was used as a reflexive verb. Most verbs (608 base lemmas from PDT) are used non-reflexively, i.e. most verbs have less than 10% of reflexive observations. The other peak is caused by mostly reflexive verbs, i.e. base lemmas with more than 90% of reflexive observations. There are only few verbs with mixed distribution. VALEVAL data confirms this observation, too.

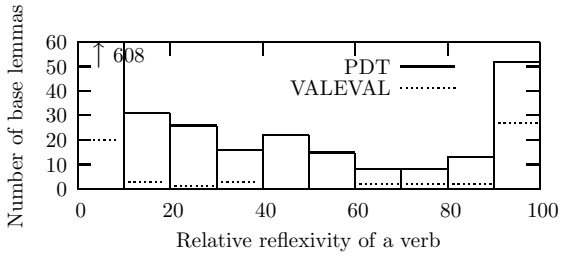

Figure 2: Relative reflexivity of base lemmas in VALEVAL and PDT.

2.4 Verb Frequencies

Figure 3 displays the number of occurrences of verbs in the Czech National Corpus (CNC[6], **?**)) compared to the number of VALLEX frames.

The chart is not based on base lemmas because by the design of VALLEX, some low frequency verbs appear in VALLEX with a high number of frames. These low frequency verbs and their frames came from a more frequent aspectual counterpart of the verb. Therefore, we merged base lemmas of both aspectual counterparts together into one "lemma cluster". The frequency of a lemma cluster is the sum of frequencies of base lemmas contained in the cluster and it ranges from a few hundreds to more than two million occurrences. The number of frames for a lemma cluster is calculated as the number of unique frames of various base lemmas.

Following the well-known experiment by Zipf, lemma clusters were sorted by descending frequency and grouped into groups of 40. The average number of frames in each group is plotted as a separate column. Zipf's law is confirmed by VALLEX and CNC data: the more frequent a verb is, the more frames it has.

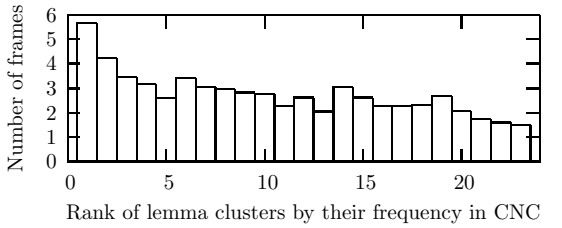

Figure 3: Occurrences of lemma clusters compared to number of VALLEX frames per lemma cluster.

3 VALEVAL: Annotating with VALLEX Frames

In order to estimate the overall quality of VALLEX and to prepare data for a frame identification or disambiguation task, we have performed a lexical sampling experiment VALEVAL[7]. For each of 109 selected base lemmas covered in VALLEX, 100 random sample sentences were extracted from CNC.

For the purposes of our experiment, verb entries were selected by these criteria: First, reflexivity of verbs was disregarded because there is no automatic procedure to identify reflexivity of a verb in a sentence (thus the sample sentences from CNC cannot be extracted separately for various reflexive variants of a verb). Second, aspectual counterparts including iteratives were grouped together to ensure that if a verb is selected, its counterparts are included as well. In order to cover both "easy" and "difficult"

[6] `http://ucnk.ff.cuni.cz/`

[7] Inspired by SENSEVAL (**?**)), a word sense disambiguation task, VALEVAL aims at valency frame disambiguation.

verbs, we ordered the groups by the total number of frames of the verb and its reflexive and aspectual counterparts. Some of the groups were selected at random in order to cover evenly the whole range of difficulty. This resulted in a more difficult annotation task for VALEVAL when compared to VALLEX in general. The average number of frames per lemma in VALLEX is 3.9 (under the label ∅) or 6.5 (labelled w∅) if weighted by number of lemma occurrences in CNC, whereas the average number of frames per lemma selected for VALEVAL is 6.8 (or 10.5 if weighted), see Table 3.

	w∅	∅
Verbs in whole VALLEX	6.47	3.92
Selected verbs	10.54	6.77

Table 3: Frames per base lemma.

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 4). Also, a special flag could be assigned to a valid answer to indicate that the annotator is not quite sure (labelled as "Uncertain 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%)
Sentences where all were sure	9280 (90.5%)
Sentences where all were sure that the frame is missing	125 (1.2%)

Table 4: Annotated data size and overall statistics about the annotations.

3.1 Inter-Annotator Agreement

Table 5 summarizes inter-annotator agreement (IAA) and Cohen's κ statistic (**?**) on the 10256 annotated sentences. The symbol ∅ indicates plain average calculated over base lemmas, w∅ stands for average weighted by frequency observed in CNC. Considering all the three parallel annotations, the exact match of answers reaches 61% (weighted) 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, as we know from Table 4).

Five different people were involved in the annotation task, the first half of annotated verbs (Verb Set 1) was annotated by three of them, the second half (Verb Set 2) was annotated by one of the first three annotators and two new ones. Apparently, the verbs in Verb Set 2 were easier to annotate, although both of the sets were selected by the same criterion: to evenly cover the whole range of verb frequencies and

complexity. The pairwise IAAs when ignoring uncertainty in both of the sets (Verb Set 1: 75.0%, 77.1%, 77.5%; Verb Set 2: 78.2%, 78.9%, 83.4%) indicate that the more the annotators are trained in FGD, the better agreement they achieve. However, the differences are not that big, if we consider the differences in annotators' familiarity with VALLEX ranging from actually authoring VALLEX to partial familiarity with FGD but seeing VALLEX for the first time. The best agreement of 83.4% was achieved by the VALLEX co-author and an expert in FGD who never used VALLEX before.

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 several different projects reported by **?**) to reach w∅ κ for noun sense disambiguation (English, French and Spanish) 0.30 to 0.49 and exceptionally up to 0.62.

	Match of 3 Annotators				**Average Pairwise Match**			
	IAA [%]		κ		IAA [%]		κ	
	w∅	∅	w∅	∅	w∅	∅	w∅	∅
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

Table 5: Inter-annotator agreement and κ.

Comparable results. Average pairwise IAA and κ measures are provided in the lower part of Table 5 to allow for a comparison with some cited results on verbs.

?) achieve an IAA for Czech verbs of 45% to 64% when annotating the Prague Dependency Treebank with Czech WordNet synsets. Their considerably lower agreement is most probably caused by relatively poor quality of the Czech WordNet.

?) reports pairwise IAA for French verbs between 60% and 65% and κ of 0.41. Our results are considerably better for both weighted and unweighted estimates, however the verb selection was done in a different manner: **?**) selected verbs of equal (moderate) frequency but at a higher level of polysemy (12.6 for the selected verbs).

?) report 90% IAA for very coarse PropBank verb framesets (based on the number of predicate-arguments only) and 71% IAA for Senseval-2 English verbs tagged with WordNet synsets, a result equal to our 70.8% for w∅ IAA. Grouping some senses together to form a more coarse grained sense inventory allowed the authors to improve the IAA to 82%. This result is comparable with a report by **?**): IAA of 86.3% for Japanese verbs. It is not clear whether the IAAs are weighted or not, nevertheless both of the figures are significantly higher than our result. This is even more surprising for the Japanese task where the average polysemy of target words was 8.3 (compared to 6.8 of ours).

Distribution of IAA and κ. Figure 4 displays the number of base lemmas in relation to the bounds of their IAA and κ. E.g., there are 26 base lemmas with IAA above 90% if the uncertainty flag is ignored. This figure rises to 34 (i.e. 31% of annotated lemmas) if we consider only sentences where all annotators were sure. We should also note that 7 base lemmas out of the total of 109 have only one frame possible.

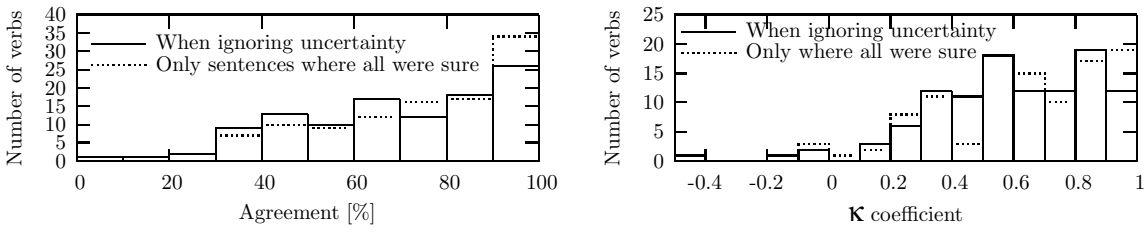

Figure 4: Histogram of IAA and κ.

The distribution of κ reveals 7 (i.e. 6%) lemmas below the threshold of 0.2 generally accepted as a *fair agreement*. Two of the four lemmas with a negative value of κ have clearly wrong VALLEX descriptions, the other two are rare aspectual counterparts with too few (< 50) occurrences in CNC for any judgements.

Based on our analysis, there is no direct relation between the IAA of κ and verb frequency or verb complexity (the number of frames defined for a given verb). The only important factor is the quality of the corresponding lexicon entries: if a verb is well structured in VALLEX, the agreement is very high, regardless the verb frequency or number of possible frames.

After carrying out a manual analysis of the annotated data of Verb Set 1 (cf. Section 6), we can conclude that the non-monotonicity of agreement when ignoring uncertainty flags in Figure 4 around 50% and 70% of agreement is mostly caused by verbs with erroneous VALLEX entries. If VALLEX frame entries had been perfect, there would have not been such gaps and we would have expected a monotone function, similar to the distribution of agreement measured on sentences where all were sure.

4 "Golden VALEVAL"

The sentences with exact agreement across the annotators (excluding invalid examples and sentences where the appropriate frame is missing) form the "golden VALEVAL".[8] Currently, the "golden VALEVAL" consists of 8066 sentences with a unique VALLEX frame assigned to the annotated verb in the sentence, see Figure 6.

	Occurrences	Lemmas
Total	8066	108
With one option only	864 (11%)	13 (12%)

Table 6: Golden VALEVAL data size.

4.1 Baselines for Frame Disambiguation

VALLEX frames correspond to verb senses (meanings). From this perspective, performing word sense disambiguation (WSD) of Czech verbs means choosing the most appropriate frame. Golden VALEVAL data can be used to evaluate automatic WSD procedures.

The difficulty of the WSD task is apparent from Table 7 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).

Based on "golden VALEVAL" sentences, Figure 5 displays the number of base lemmas with the relative frequency of the most frequent frame. As we see, the "assign most frequent" strategy hurts some base lemmas more and some less. **?**) predicts that the dominance of the commonest frame (i.e. the baseline accuracy when assigning most frequent frame) should rise for more frequent verbs. Eg., the verbs contributing to the column at 90% should be in general more frequent than the verbs contributing to the column at 50%. VALEVAL data is not representative enough to support this prediction.

[8] Moreover, the sentences for the first half of lemmas were cross-checked and some of them were supplied with an authoritative answer eliminating typing errors and clear misinterpretations in order to extend the "golden VALEVAL" collection.

	w∅	∅
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

Table 7: Baselines for WSD.

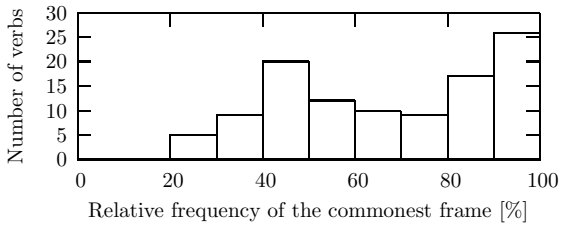

Figure 5: The number of base lemmas where the commonest frame has the appropriate relative frequency in "golden VALEVAL" sentences.

5 Frame Disambiguation Experiment

As input for the frame disambiguation task we take VALEVAL corpus described in the previous section. First, we automatically parse the data using the Charniak parser (**?**). After excluding unparsed sentences, 6666 sentences for 107 different verbs remain, counting 62.3 sentences for verb on average.

We employed the machine learning toolkit C5.0[9] which implements two methods: **decisions trees** and **sets of if-then rules**. Again, the two machine learning methods were evaluated using 10-fold cross-validation.

5.1 Features Used

For automatic determination of VALLEX frame to which a given verb belongs, we generate a vector of features for each verb occurred in the data.

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 words in a small window containing words neighboring the verb in the surface representation of the sentence.
- Syntax-based: information dependent on the result of the automatic syntactic parser (including mainly morphologic and lexicographic information).
- Idiomatic: occurrence of phasems in the sentence according to the VALLEX lexicon.

5.1.1 Morphological Features

Morphological sign for Czech consists of 15 positions where each position states value of a given category (e.g. part of speech, number, gender, ...). Categories which are not relevant for a given word form (e.g. tense for nouns) have a special value ("–").

For five words around the verb occurrence (two preceding words, the verb itself, and two following words) we take each position of each morphological tag as a single feature, counting 75 morphologic features together (5 words, 15 featured each).

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

	w∅		∅	
Type of feature	**Decision Trees**	**If-then rules**	**Decision Trees**	**If-then rules**
Baseline	63.31 %		67.94 %	
Morphological (M)	69.22 %	68.03 %	73.86 %	73.73 %
Syntax-based (S)	72.26 %	72.55 %	78.58 %	78.6 %
Idiomatic (I)	63.50 %	63.53 %	67.97 %	67.99 %
M + S	70.06 %	70.87 %	79.16 %	79.04 %
M + I	67.69 %	68.56 %	73.73 %	73.94 %
S + I	72.53 %	72.38 %	78.69 %	78.59 %
M + S + I	69.34 %	70.78 %	79.11 %	79.03 %

Table 8: Accuracy of frame selection using different types of features.

5.1.2 Syntax-based Features

Based on the dependency tree (constructed automatically by the parser) we extracted the following boolean features for each verb occurrence:

- **reflexive particles**: two features denoting the presence of reflexive particles "se" and "si" dependent on the verb
- **superordinate verb**: a feature denoting whether the verb depends on another verb
- **subordinate verb**: a feature denoting presence of a subordinate verb dependent on the verb
- **subordinate conjunctions**: six features denoting presence of subordinate conjunctions (*aby*, *at'*, *až*, *jak*, *že*, *zda*) dependent on the verb
- **nouns in cases**: seven features, one for each case denoting presence of a noun or a substantive pronoun in the given case dependent on the verb
- **adjectives in cases**: seven features, one for each case denoting presence of an adjective or an adjectival pronoun in the given case dependent on the verb
- **prepositional phrase in case**: seven features, one for each case denoting presence of a prepositional phrase in the given case dependent on the verb
- **degree**: three features denoting presence of a child in first, second and third degree respectively
- **prepositions in cases**: 69 features, one for each possible combination of a preposition and a case, denoting the presence of a preposition in the given case dependent on the verb

5.1.3 Idiomatic Features

We extracted one boolean feature for each idiomatic expression defined in the VALLEX lexicon. We set the value of the corresponding feature to **true** if the idiom occurs anywhere in the sentence (the word forms of the idiomatic expression occur in a row in the surface representation of the sentence).

5.2 Results

Table 8 states results of decision trees and rule-based learning taking each group of features separately and in combination. Again, the accuracy figures are given weighted (w∅) by the number of occurrences of each verb in CNC and unweighted (∅). Note that the figures have to be compared to the baseline given here in Table 8, which is slightly higher than the "Golden VALEVAL" as a whole due to the removal of unparsed sentences.

From the results it is obvious that adding information about idiomatic expresions did not bring any improvement at all. This is mainly due to the low number of sentences with idiomatic expressions in the data (together there were 323 idioms in the data, i.e. 3 idioms for a verb on average).

Feature type	Feature	Weight
Syntax-based	Reflexive particle *se* dependent on the verb	52
Syntax-based	Preposition in acusative dependent on the verb	22.5
Morphological	Detailed part of speech of the word two positions after the verb	18
Morphological	Case of the word two positions after the verb	14
Syntax-based	Noun or nominative pronoun in dative dependent on the verb	12.5
Morphological	Detailed part of speech of the word preceeding the verb	12
Syntax-based	Preposition in dative dependent on the verb	11.3
Morphological	Gender of the word following the verb	8
Syntax-based	Preposition *z* in genitive dependent on the verb	7.1
Syntax-based	An infinitive verb dependent on the verb	7
Morphological	Voice of the verb	6.1
Morphological	Number of the word following the verb	6
Morphological	Number of the verb	5.5
Syntax-based	Preposition in genitive dependent on the verb	5.5
Morphological	Gender of the verb	5
Morphological	Case of the word following the verb	5
Syntax-based	Reflexive particle *si* dependent on the verb	5

Table 9: Features with the biggest overall influence.

Adding morphological features to syntax-based features also did not bring any substantial improvement. This supports an idea that the local syntactic neighbourhood of the verb depicts enough information to achieve the rate of disambiguation which can be obtained using this method. In the case of weighting the accuracy by the frequencies in the CNC, adding morphological features to syntax-based featyres led to even worse results. We think that this is mainly due to sparse data we are using, and the sparse data problem is more severe for highly frequented verbs.

5.3 Most Important Features

Table 9 shows which features have the biggest influence on the WSD task using decision trees. The learning algorithm builds a decision tree for each verb. Each node of the tree represents a decision point based on the value of a single feature. The closer to the root, the more important the feature is with respect to the task of selecting the correct verb frame. Therefore, we weight the features with the 0.5-exponent of the step at which it is used (features used at the roots of the trees get weight of 1.0, features used at the second step get weight of 0.5, at the third step, weight of 0.25 is used, etc.). For each feature we sum up the weights obtained in trees for all the verbs.

The automatically learned decision trees confirm basic linguistic expectations: for the choice of the verb frame, the most important is the presence of a reflexive particle (cf. section 2.3). The second most important test deals with intransitivity of the verb (ie. the presence of an object in accusative and the respective preposition used). A test on ditransitivity comes next, expressed in a number of features: the secondary object appears usually two positions after the verb as a noun or pronoun in dative.

6 VALLEX 1.0 Corrections

After a manual analysis of results of annotating Verb Set 1 (the first half of data), we propose 137 corrections to VALLEX data. The corrections proposed are of the following types:

- frame entries (75 corrections in total)

  - missing frame entries (57 corrections) *Táhla se s taškami domů.(She lugged the bags home.) Chabá koruna může obrátit vývoj obchodní bilance. (Easy crown can turn the development of the balance of trade.)*
    High number of missing valency frames is caused by high amount of missing idioms (39 verb entries from total of 57 missing frame entries). VALLEX 1.0 does not intend to cover all verb idioms. It records only the most frequent ones. For illustration, the number of idioms of verbs with rich polysemy such as *brát, dát, jít* (take, give, go etc.) is according to **?**) approximately hundred. A similar observation is given by **?**) for English. We consider missing verb entries for non-idioms as a more serious problem.
  - inappropriately joined or split frame entries (6 corrections)
    Přidala sůl do jídla. Přidal k lustracím prohlášení. (She added salt to the food. He supplied the documents with a declaration.)
    The previous examples are covered by one frame entry ACT(1;obl) PAT(4;obl) DIR3(;obl) with the following gloss: *dodat, připojit* (supply, add). We consider dividing these examples into two frame entries. Thus, a new frame entry with the valency frame ACT(1;obl) PAT(4;obl) EFF(k+3;obl) is to be added.
  - superfluous frame entries (12 corrections)
    Spojil obyvatele do sdružení. (He joined/partnered inhabitants into an association.) Spojil procházku s nákupem. (He joined a walk with shopping.)
    Here, the two separate frame entries ought to be merged.
- mistakes within frame entries (32)
  - mistakes in functors (16)
    - inappropriately chosen functor (12)
      Zůstal bez peněz. He remains out of pocket.
      ACT(1;obl) ACMP(bez+2;obl) → (should be changed to) ACT(1;obl) PAT(bez+2;obl)
    - missing functor (4)
      Situace se obrátila v katastrofu. The situation turned into the catastrophe.
      ACT(1;obl) → ACT(1;obl) PAT(v+4;opt)
  - mistakes in morphemic realization (12)
    - inappropriately chosen form (1)
      Nechal dveře otevřené. (He left door open.)
      ACT(1;obl) PAT(4;obl) EFF(adj-1;obl) → ACT(1;obl) PAT(4;obl) EFF(adj-4;obl)
    - missing form (11)
      Rozdělil lidi do skupin / na skupiny / ve skupiny. (He divided the people into the groups.)
      ACT(1;obl) PAT(4;obl) EFF(do+2,na+4;opt) → ACT(1;obl) PAT(4;obl) EFF(do+2,na+4,v+4)
  - wrong type of complementations (4)
    - change from obligatory to optional
      Soustředí se na práci. Soustředí se. (He is concentrating on work. He is concentrating.)
      ACT(1;obl) PAT(na+4;obl) → ACT(1;obl) PAT(na+4;opt)
    - change from obligatory to typical (2)
      Povolil mu ohledně té věci. (He compromised with him regarding that matter / thing.)
      ACT(1;obl) PAT(3;opt) REG(uhledně+2,v+6;obl) → ACT(1;obl) PAT(3;opt) REG(ohledně+2, v+2;typ)
    - change from typical into obligatory
      Orientovali okna na jih. (They oriented the windows to the south.)
      ACT(1;obl(PAT(4;obl) NORM(podle+2;typ) DIR3(;typ) → ACT(1;obl) PAT(4;obl) DIR3(;obl) NORM(podle+2;typ)
- mistakes in the attributes of verb entries (30 corrections)
  - gloss (23)

    - supply the gloss with a missing meaning of the verb (13)
      Řešení odpovídá svými vlastnostmi požadavkům. (The solution corresponds to the requirements by its features.)
      gloss: být ve shodě → být ve shodě / v souladu / korespondovat
      gloss: be in accordance with / be in compliance with → be in accordance with / be in compliance with / correspond with
    - inappropriate gloss restricts the meaning (6)
      Přidal synovi na auto. (He contributed to a car.)
      gloss: přispět → přispět; dát
      gloss: contribute → contribute / give
    - inappropriate gloss broadens vaguely the meaning of the verb (4)
      Pole se táhla až k lesu. (The fields extend to the forrest.)
      gloss: být dlouhý → rozprostírat se
      gloss: be long → extend / be long
  - inappropriate examples (7)

We treat superfluous verb entries, missing verb entries (except for the missing idioms), wrong type of complementations, incorrect choice of functor or its morphemic realization as a more serious problems than the missing verb entries for idioms, missing morphemic realization, mistakes in glosses and examples. To conclude, approximately 68 corrections of serious mistakes need to be carried out in VALLEX 1.0.

7 Conclusion

We described VALEVAL, an annotation experiment using a sample of 10256 sentences with VALLEX frames. Inter-annotator agreement of three parallel annotations ranges from 61% to 74% and κ from 0.52 to 0.62 relative to the type of exact match required. The key factor determining the agreement is the quality of lexicon entries, not the overall frequency of the verb or the number of possible frames.

More than 8000 sentences are currently available as "golden VALEVAL" corpus for WSD experiments with a baseline accuracy of 60% (weighted by verb frequencies) or 66% (unweighted). We are now incorporating "golden VALEVAL" to the Senseval-4[10] task. We also conducted some experiments ourselves and the accuracy achieved by our method reaches 70% (weighted) or 80% (unweighted).

In addition, VALEVAL allowed us to check the consistency of VALLEX 1.0 and nearly 70 serious mistakes in VALLEX 1.0 were found. VALEVAL provided us with an important feedback for further improvements of VALLEX.

Acknowledgement

We are grateful to Markéta Lopatková, Kiril Ribarov and Zdeněk Žabokrtský for many insightful comments on this paper. This work has been partially supported by the grants GAČR No. 201/05/H014 (O. Bojar), GA UK No. 372/2005/A-INF/MFF (J. Semecký) and Program "Information Society" under project 1ET100300517 (V. Benešová).

[10] http://www.senseval.org/

References

Babko-Malaya, Olga, Martha Palmer, Nianwen Xue, Aravind Joshi, and Seth Kulick. 2004. Proposition Bank II: Delving Deeper. In *Proc. of the Frontiers in Corpus Annotation Workshop 2004, in conjunction with NAACL-HLT'04*.

Carletta, Jean. 1996. Assessing agreement on classification task: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Čermák, František, Jiří Hronek, et al., editors. 1994. *Slovník české frazeologie a idiomatiky*. Academia, Praha.

Charniak, Eugene. 2000. A Maximum-Entropy-Inspired Parser. In *Proceedings of NAACL-2000*, pages 132–139, April.

Chklovski, Timothy and Rada Mihalcea. 2003. Exploiting Agreement and Disagreement of Human Annotators for Word Sense Disambiguation. In *Proceedings of Recent Advances In NLP (RANLP 2003)*, September.

Edmonds, Philip. 2002. Introduction to senseval. *ELRA Newsletter*, October 2002.

Fillmore, Charles J. 2002. FrameNet and the Linking between Semantic and Syntactic Relations. In Shu-Cuan Tseng, editor, *Proceedings of COLING 2002*, pages xxviii–xxxvi. Howard International House.

Hajič, Jan, Martin Holub, Marie Hučínová, Martin Pavlík, Pavel Pecina, Pavel Straňák, and Pavel Šidák. 2004. Validating and Improving the Czech WordNet via Lexico-Semantic Annotation of the Prague Dependency Treebank. In *Proceedings of LREC 2004*.

Hajič, Jan, Jarmila Panevová, Eva Buráňová, Zdeňka Urešová, and Alla Bémová. 2001. A Manual for Analytic Layer Tagging of the Prague Dependency Treebank. Technical Report TR-2001-, ÚFAL MFF UK, Prague, Czech Republic. English translation of the original Czech version.

Hajič, Jan, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová, and Petr Pajas. 2003. 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*, pages 57–68. Vaxjo University Press, November 14–15, 2003.

Kilgarriff, Adam. 2004. How Dominant is the Commonest Sense of a Word? Technical Report ITRI-04-10, Information Technology Research Institute, University of Brighton. Also published in Proceedings of TSD 2004, Text, Speech and Dialogue 7th International Conference, Brno, Czech Republic, September 2004.

Kingsbury, Paul, Martha Palmer, and Mitch Marcus. 2002. Adding Semantic Annotation to the Penn TreeBank. In *Proceedings of the Human Language Technology Conference*.

Kocek, Jan, Marie Kopřivová, and Karel Kučera, editors. 2000. *Český národní korpus - úvod a příručka uživatele*. FF UK - ÚČNK, Praha.

Lopatková, Markéta and Jarmila Panevová. 2004. Recent developments of the theory of valency in the light of the Prague Dependency Treebank. *SNK*. in press.

Panevová, Jarmila. 1994. Valency Frames and the Meaning of the Sentence. In Ph. L. Luelsdorff, editor, *The Prague School of Structural and Functional Linguistics*, pages 223–243, Amsterdam-Philadelphia. John Benjamins.

Pustejovsky, James, Patrick Hanks, and Anna Rumshisky. 2004. Automated Induction of Sense in Context. In Silvia Hansen-Schirra, Stephan Oepen, and Hans Uszkoreit, editors, *COLING 2004 5th International Workshop on Linguistically Interpreted Corpora*, pages 55–58, Aug 29.

Sgall, Petr, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecht, Netherlands.

Shirai, Kiyoaki. 2002. Construction of a Word Sense Tagged Corpus for SENSEVAL-2 Japanese Dictionary Task. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 605–608, May.

Straňáková-Lopatková, Markéta and Zdeněk Žabokrtský. 2002. Valency Dictionary of Czech Verbs: Complex Tectogrammatical Annotation. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, volume 3, pages 949–956. ELRA.

Véronis, Jean. 1998. A study of polysemy judgements and inter-annotator agreement. In *Programme and advanced papers of the Senseval workshop*, pages 2–4, Herstmonceux Castle (England).

Žabokrtský, Zdeněk. 2005. *Valency Lexicon of Czech Verbs*. Ph.D. thesis, Faculty of Mathematics and Physics, Charles University in Prague. in prep.

Žabokrtský, Zdeněk and Markéta Lopatková. 2004. Valency Frames of Czech Verbs in VALLEX 1.0. In *Frontiers in Corpus Annotation. Proceedings of the Workshop of the HLT/NAACL Conference*, pages 70–77, May 6, 2004.