Czech Verbs of Communication and the Extraction of their Frames*

Václava Benešová and Ondřej Bojar

Institute of Formal and Applied Linguistics
ÚFAL MFF UK, Malostranské náměstí 25, 11800 Praha, Czech Republic
{benesova,bojar}@ufal.mff.cuni.cz

Abstract. We aim at a procedure of automatic generation of valency frames for verbs not covered in VALLEX, a lexicon of Czech verbs. We exploit the classification of verbs into syntactico-semantic classes. This article describes our first step to automatically identify verbs of communication and to assign the prototypical frame to them. The method of identification is evaluated against two versions of VALLEX and FrameNet 1.2. For the purpose of frame generation, a new metric based on the notion of frame edit distance is outlined.

1 Introduction

The main objective of this paper is to present first experiments with an automatic extension of VALLEX, a valency lexicon of Czech verbs. Czech verbs were included in the lexicon on the basis of their frequency in the Czech National Corpus (CNC[1]) to achieve maximal corpus coverage. VALLEX nowadays covers around 66% of verb occurrences; 23% of verb occurrences belong to few frequent auxiliary verbs, esp. *být, bývat (to be)*. (See Table 1.) The remaining 10% occurrences belong to verbs with low corpus frequency. It would not be economical to continue manual development of VALLEX for the remaining entries because the distribution of verbs closely follows Zipf's law and there are about 28k verbs needed just to cover our particular corpus.

In order to cover the missing verbs, we have experimented with the possibility of automatic generation of frames based on corpus evidence. Our experiment exploits the classification of verbs into semantic classes, a piece of information which is already available in VALLEX. For the time being, we have focused on a single class: verbs of communication, the so called *verba dicendi*.

In our contribution, we first provide a basic description of VALLEX 1.x, including its classification of verbs. The examined class of verbs of communication is described in a greater detail. Next, we describe and evaluate the proposed automatic method to identify verbs of communication. Finally, we estimate the usefulness of the identification of this class in the task of automatic creation of VALLEX entries.

* The work reported in this paper has been supported by the grants GAAV ČR 1ET201120505, LC536 and GAČR No. 405/04/0243.

[1] http://ucnk.ff.cuni.cz/

Table 1. Coverage of VALLEX 1.0 and 1.5 with respect to the Czech National Corpus.

	VALLEX 1.0				VALLEX 1.5			
	Occ.	[%]	Verb lemmas	[%]	Occ.	[%]	Verb lemmas	[%]
Covered	8.0M	53.7	1,064	3.6	8.0M	65.6	1,802	6.1
Not covered but frequent	4.1M	27.9	20	0.1	3.5M	23.4	4	0.0
Not covered, infrequent	2.7M	18.3	28,385	96.3	1.6M	10.9	27,663	93.9
Total	14.8M	100.0	29,469	100.0	14.8M	100.0	29,469	100.0

1.1 VALLEX, Valency Lexicon of Czech Verbs

VALLEX uses the Functional Generative Description [1] as its theoretical background and is closely related to the Prague Dependency Treebank (PDT, [2]). VALLEX is fully manually annotated, which sets limits on the growth rate. On the other hand, manual annotation ensures attaining data of high quality. The first version of VALLEX 1.0 was publicly released in 2003 and contained over 1,400 verb entries[2]. The set of covered verbs was extended to about 2,500 verb entries in VALLEX 1.5, an internal version released in 2005. (See also Table 2.) The second version, VALLEX 2.0 (almost 4,300 entries) based on the so-called alternation model (see [3]), will be available in autumn 2006.

VALLEX 1.0 and 1.5 consist of verb entries containing a non-empty set of valency frames. Under the term *valency*, we understand the ability of a verb to bind a range of syntactic elements. A valency frame is assigned to a verb in its particular meaning/sense and is captured as a sequence of frame slots. Each slot stands for one complement and consists of a functor (a label expressing the relation between the verb and the complement), its morphemic realization and the type of complement.

1.2 Verb Classes in VALLEX

Verb classes were introduced to VALLEX primarily to improve data consistence because observing whole groups of semantically similar verbs together simplifies data checking.

At present, classification of verbs into semantic classes is a topical issue in linguistic research (cf. Levin's verb classes [4], PropBank [5] , LCS [6, 7], FrameNet [8]). Although we consider these approaches to be very stimulative from the theoretical point of view, we decided to use our own classification for the reason of differences in the theoretical background and in the methods of description.

However, we must emphasize that building verb classes and their description in VALLEX is still in progress and the classification is not based on a defined

[2] The term *verb entry* refers to a VALLEX entry which distinguishes homographs and reflexive variants of the verb. The term *verb lemma* refers to the infinitive form of the verb, excluding the reflexive particle.

ontology but is to a certain extent intuitive. VALLEX classes are built thoroughly from below. When grouping verbs together, we give priority mostly to syntactic criteria: the number of complements (FGD classifies them into inner participants, the so-called *actants*, and *free modifications* roughly corresponding to adjuncts), their type (mainly obligatory or optional), functors and their morphemic realizations.

As displayed in Table 2, VALLEX now defines about 20 verb classes (communication, mental action, perception, psych verbs, exchange, change, phase verbs, phase of action, modal verbs, motion, transport, location, expansion, combining, social interaction, providing, appoint verb, contact, emission, extent) that contain on average 6.1 distinct frame types (disregarding morphemic realizations and complement types).

Table 2. Basic statistics about VALLEX 1.0 and 1.5.

	VALLEX 1.0	VALLEX 1.5
Total verb entries	1,437	2,476
Total verb lemmas	1,081	1,844
Total frames	4,239	7,080
Frames with a class	1,591 (37.5%)	3,156 (44.6%)
Total classes	16	23
Avg. frame types in class	6.1	6.1

1.3 Verbs of Communication

The communication class is specified as the set of verbs that render a situation when 'a speaker conveys information to a recipient'. Besides the slots ACT for the 'speaker' and ADDR for the 'recipient', communication verbs are characterized by the entity 'information' that is expressed on the layer of surface structure as a dependent clause introduced by a subordinating conjunction or as a nominal structure.

On the one hand, the entity 'information' is the property that relates these verbs to verbs of some other classes (mental action, perception and psych verbs). On the other hand, the inherence of the 'recipient' distinguishes the verbs of communication from the aforementioned other classes. However, in a small number of cases when the addressee which represents the 'recipient' does not appear explicitly in valency frame (*speak, declare, etc.*), this distinctive criterion fails.

On the basis of our observations, the verbs of communication can be further divided into subclasses according to the semantic character of 'information' as follows: simple information (verbs of announcement: *říci (say), informovat (inform)*, etc.), questions (interrogative verbs: *ptát se (ask), etc.*) and commands, bans, warnings, permissions and suggestions (imperative verbs: *poručit (order),*

zakázat (prohibit), etc.).The dependent clause after verbs of announcement is primarily introduced by the subordinating conjunction *že (that)*, interrogative by *zda (whether), jestli (if)* and imperative verbs by *aby (in order to), ať (let)*. We recognize some other distinctions between these three subclasses but their description goes beyond the scope of this paper.

2 Automatic Identification of Verbs of Communication

In the present section, we investigate how much the information about valency frame combined with the information about morphemic realization of valency complement can contribute to an automatic recognition of verbs of communication. For the sake of simplicity, we use the term *verbs of communication* to refer to verbs with at least one sense (frame) belonging to the communication class.

2.1 Searching Corpus for Typical Surface Patterns

Our experiment is primarily based on the idea that verbs of communication can be detected by the presence of a dependent clause representing the 'information' and an addressee representing the 'recipient'.

This idea can be formalized as a set of queries to search the corpus for occurrences of verbs accompanied by: (1) a noun in one of the following cases: genitive, dative and accusative (to approximate the ADDR slot) and (2) a dependent clause introduced by one of the set of characteristic subordinating conjunction (*že*, *aby*, *ať*, *zda* or *jestli*) (to approximate the slot of 'information').

We disregard the freedom of Czech word order which, roughly speaking, allows for any permutation of a verb and its complements. In reality, the distribution of the various reorderings is again Zipfian with the most typical pattern (verb+ADDR+subord) being the most frequent. In a sense, we approximate the sum with the first, maximal, element only. On the other hand we allow some intervening adjuncts between the noun and the subordinating clause.

2.2 Evaluation against VALLEX and FrameNet

We sort all verbs by the descending number of occurrences of the tested pattern. This gives us a ranking of verbs according to their 'communicative character', typical verbs of communication such as *říci* (say) appear on top. Given a threshold, one can estimate the class identification quality in terms of a confusion matrix: verbs above the threshold that actually belong to the class of verbs of communication (according to a golden standard) constitute *true positives* (TP), verbs above the threshold but not in the communication class constitute *false positives* (FP), etc.

A well-established technique of the so-called ROC curves allows to compare the quality of rankings for all possible thresholds at once. We plot the *true positive rate* ($TPR = TP/P$ where P is the total number of verbs of communication)

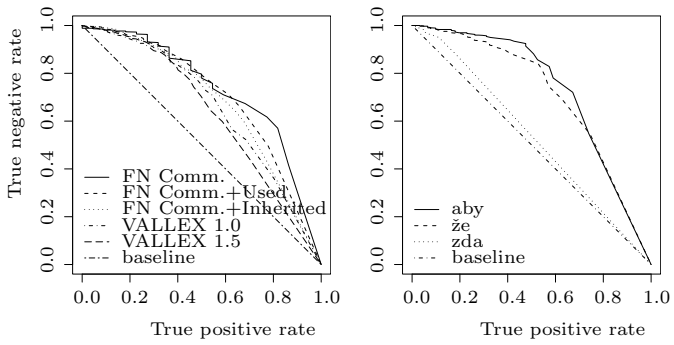

Fig. 1. Verbs of communication as suggested by the pattern V+N234+subord, evaluated against VALLEX and FrameNet (left) and evaluated against VALLEX 1.0 for three main contributing subordinating conjunctions (*aby*, *že*, *zda*) independently (right).

against *true negative rate* ($TNR = TN/N$, N stands for the number of verbs with no sense of communication) for all thresholds.

We evaluate the quality of class identification against golden standards from two sources. First, we consider all verbs with at least one frame in the communication class from VALLEX 1.0 and 1.5 and second, we use all possible word-to-word translations of English verbs listed in FrameNet 1.2[3] Communication frame and all inherited and used frames (For an explanation, see [9, 10]; the English-to-Czech translations were obtained automatically using available on-line dictionaries). As the universum (i.e. $P + N$), we use all verbs defined in the respective version of VALLEX and all verbs defined in VALLEX 1.5 for the FrameNet-based evaluation.

Figure 1 displays the TPR/TNR curve for verbs suggested by the pattern V+N234+subord. The left chart compares the performance against various golden standards, the right chart gives a closer detail on contribution from different subordinating conjunctions.

The closer the curve lies to the upper right corner, the better the performance is compared to the golden standard. With an appropriate threshold, about 40% to 50% of verbs of communication are identified correctly while 20% of non-communication verbs are falsely marked, too. We get about the same performance level for both VALLEX and FrameNet-based evaluation. This confirms that our method is not too tightly tailored to the classification introduced in VALLEX.

The right chart in Figure 1 demonstrates that the contribution of different subordinating conjunctions is highly varied. While *aby* and *že* contribute significantly to the required specification, the verbs suggested by the pattern with *zda* are just above the baseline of not suggesting any verb. (The conjunctions *ať* and *jestli* had too few occurrences in the pattern.)

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

2.3 Weak Points of Patterns

From the very beginning, we eliminated the nominal structures (which can also express 'information') from the queries in order to avoid verbs of exchange as *give*, *take*, etc. In a similar vein, the queries were not able to identify sentences with verbs of communication where some of the searched complements were not realized on the layer of surface structure. Therefore, some verbs which belong to the communication class remained undiscovered.

On the contrary, the fact that conjunctions *aby* and *že* are homonymous lowers the reliability of the queries. We tried to eliminate the number of incorrectly chosen verbs by a refinement of the queries. (For instance, we omitted certain combination of demonstratives plus conjunctions: *tak*, *aby* (*so that*), *tak*, *že* (*so that*), etc.) A further problem is represented by cases when the identified dependent clause is not a member of the valency frame of the given verb but depends on the preceding noun.

3 Frame Suggestion

One of our foreseen tasks is to generate VALLEX frame entries for new verbs based on corpus data. This is a well-established research topic (see [11] for a survey) but most experiments were conducted with focus on surface frames only, making the experimental setting comparably easier.

The method of searching corpus for typical patterns described in the previous section can contribute to frame extraction task in the following manner: for all verbs occurring frequently enough in the typical pattern, we propose the most typical 'communication frame' consisting of ACT, ADDR and PAT (all obligatory). For each verb independently, we assign only conjunctions discovered by the queries to the PAT. Every verb of communication can have some additional senses not noticed by our method but at least the communication frame should be suggested correctly.

3.1 Frame Edit Distance and Verb Entry Similarity

Methods of frame extraction are usually evaluated in terms of precision and recall of either frames as wholes or of individual frame elements (slots). These metrics are unfortunately too rough for the richly structured VALLEX-like frames. Therefore, we propose a novel metric, *frame edit distance* (FED). The metric estimates the number of edit operations (insert, delete, replace) necessary to convert a hypothesized frame to a correct frame. In the current simple version of the metric, we assign equal costs to all basic edit operations (fixing the obligatoriness flag, adding or removing allowed morphemic forms), only the functor is considered as fixed. In order to change the functor, one pays for complete destruction of the wrong slot and complete construction of the correct slot. We consider charging more for slot destruction than for slot construction in future versions of the metric because we prefer methods that undergenerate and produce safer frames to methods that suggest unjustified frames.

As described above, VALLEX is organized as a set of verb entries each consisting of a set of frames. Given a verb lemma, the set of its VALLEX entries and a set of entries produced by an automatic frame suggestion method, we can use FED to estimate how much of editing work has been saved. We call this measure *entry similarity* or *expected saving* (ES) and define it as follows:

$$ES = 1 - \frac{\min FED(G,H)}{FED(G,\emptyset)+FED(H,\emptyset)}$$

G denotes the set golden verb entries of this base lemma, H denotes the hypothesized entries and $\emptyset$ stands for a blank verb entry. Not suggesting anything has ES of 0% and suggesting the golden frames exactly has ES of 100%.

3.2 Experimental Results with Verb Entry Similarity

Table 3 displays the ES of four various baselines and the result obtained by our method. When we assume that every verb has a single entry and this entry consists of a single frame with the ACT slot only, ES estimates that about 27% of editing operations was saved. Suggesting ACT and PAT helps even better (Baseline 2, 38%), but suggesting a third obligatory slot for ADDR (realized either as dative (3) or accusative (4)) is already harmful, because not all the verb entries require an ADDR.

We can slightly improve over Baseline 2 if we first identify verbs of communication automatically and assign ACT PAT ADDR with appropriate subordinating conjunctions to them, leaving other verbs with ACT PAT only. This confirms our assumption that verbs of communication have a typical three-slot frame and also that our method managed to identify the verbs correctly.

Our ES scores are relatively low in general and Baseline 4 suggests a reason for that: most verbs listed in VALLEX have several senses and thus several frames. In this first experiment, we focus on the communication frame only, so it still remains quite expensive (in terms of ES) to add all other frames. In Baseline 4, we suggest a single verb entry with two core frames (ACT PAT) and this gives us a higher saving because most verbs indeed ask for more frames.

Table 3. Expected saving when suggesting frame entries automatically.

Suggested frames	ES [%]
Specific frame for verbs of communication, default for others	38.00 ± 0.19
Baseline 1: ACT(1)	26.69 ± 0.14
Baseline 2: ACT(1) PAT(4)	37.55 ± 0.18
Baseline 3: ACT(1) PAT(4) ADDR(3,4)	35.70 ± 0.17
Baseline 4: Two identical frames: ACT(1) PAT(4)	39.11 ± 0.12

4 Conclusion

We briefly described the classification of verbs in VALLEX and we proposed and evaluated a corpus-based automatic method to identify verbs of communication. The performance of our method was tested not only on VALLEX data but also on an independent verb classification as available in the FrameNet.

We introduced a novel metric to capture the effort to construct VALLEX verb entries and to estimate how much effort an automatic procedure can save. Having assigned a prototypical frame of communication to the verbs that were automatically identified in the previous step, we achieved a little improvement over the baseline, although not statistically significant.

We conclude that the automatic identification of communication verbs proposed performs satisfactorily. However, to employ this step in an automatic generation of verb entries for new verbs, the method must not be restricted to a single class and suggest also other frames for other verb senses. Otherwise, only very little of lexicographic labour is saved.

References

1. Sgall, P., Hajičová, E., Panevová, J.: The Meaning of the Sentence in Its Semantic and Pragmatic Aspects. D. Reidel Publishing Company, Dordrecht (1986)
2. Hajič, J.: Complex Corpus Annotation: The Prague Dependency Treebank. In Šimková, M., ed.: Insight into Slovak and Czech Corpus Linguistics, Bratislava, Slovakia, Veda, vydavateľstvo SAV (2005) 54–73
3. Lopatková, M., Žabokrstký, Z., Skwarska, K.: Valency Lexicon of Czech Verbs: Alternation-Based Model. In: Proceedings of LREC 2006, ELRA (2006) 1728–1733
4. Levin, B.: English Verb Classes and Alternations. University of Chicago Press, Chicago (1993)
5. Palmer, M.e.a.: The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics **31**(1) (2005) 71–106
6. Jackendoff, R.: Semantic Structures. The MIT Press, Cambridge, MA (1990)
7. Dorr, B.J., Mari, O.B.: Multilingual Generation: The Role of Telicity in Lexical Choice and Syntactic Realization. Machine Translation **11**(1–3) (1996) 37–74
8. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The Berkeley FrameNet project. In Boitet, C., Whitelock, P., eds.: Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics, San Francisco, California, Morgan Kaufmann Publishers (1998) 86–90
9. Fillmore, C.J., Wooters, C., Baker, C.F.: Building a large lexical databank which provides deep semantics. In: Proceedings of the Pacific Asian Conference on Language, Information and Computation, Hong Kong (2001)
10. Fillmore, C.J.: FrameNet and the Linking between Semantic and Syntactic Relations. In Tseng, S.C., ed.: Proceedings of COLING 2002, Howard International House (2002) xxviii–xxxvi
11. Korhonen, A.: Subcategorization Acquisition. Technical Report UCAM-CL-TR-530, University of Cambridge, Computer Laboratory, Cambridge, UK (2002)