Measuring Coverage of a Valency Lexicon using Full Syntactic Analysis

Miloš Jakubíček, Vojtěch Kovář, and Aleš Horák

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{xjakub,xkovar3,hales}@fi.muni.cz

Abstract. Recent development showed that valency information provides a great benefit in many areas of natural language processing. Building valency lexicons is however a complex and time-consuming task from both theoretical and practical points of view, since designing of the lexicon plays a crucial role in its future usability as well as its careful and considerated preparation. As for any manually created resource, it is complicated to evaluate its quality. In this paper we consider the usage of the syntactic parser `synt` for estimating the coverage of the Verbalex verb valency lexicon for Czech. For this task we extended the phrase extraction functionality of the parser, which we describe briefly. Finally we discuss our results and further development.

Key words: verb valency; syntactic analysis; lexicon coverage

1 Introduction

During the last decade researchers tried to enhance their NLP applications by supplying additional linguistic information. The usage of all kinds of resources from basic lexicons over annotated corpora to complex ontologies has proved to be necessary for further development of most computational linguistics applications. Their preparation is however very time-consuming and therefore costly since the prevailing majority of them needs to be created manually (at least partially). This raises many problems in both design and implementation of a particular resource: it ought to be carefully designed from the theoretical point of view because the opportunities to modify it automatically in the future are limited. The actual preparation also has to be addressed attentively to ensure consistency, validity and completeness of the prepared data.

It is therefore of great benefit to make any of these steps a bit easier, e. g. to move from fully manual to semi-automatic processing and to provide (semi-)automatic evaluation methods. In this paper we describe one of such steps that we have taken in the case of building the Verbalex verb valency lexicon for Czech [1]. We used a Czech parser called `synt` [2] for automatic extraction of shallow verb valencies from the DESAM corpus [3] which is manually annotated. The proposed method gives an estimation of the lexicon coverage as well as speeds up the preparation of the lexicon by providing preprocessed data for annotators and offer suitable examples for existing verb valencies.

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2009, pp. 75–79, 2009.

2 Syntactic Parser `synt`

The syntactic parser `synt` [4] has been developed for several years in the Natural Language Processing Centre at Masaryk University. It performs an agenda-based head-corner chart parsing using the provided context-free grammar for Czech. For easy maintenance this grammar is recorded in the form of a metagrammar (having about 200 rules) from which the full grammar can be automatically derived (having almost 4,000 rules). Contextual phenomena (such as case-number-gender agreement) are covered using the contextual actions defined for each rule.

It has been shown that `synt` achieves a very good coverage (more than 90 % [5, p. 77]), but the analysis it provides is highly ambiguous: for some sentences even millions of output syntactic trees can occur. There are two main strategies developed to fight such ambiguity. First, the grammar rules are divided into different priority levels that are used to prune the resulting set of output trees. Second, each grammar rule has a ranking value assigned from which the ranking for the whole tree can be efficiently computed in order to select only the best trees for the output.

This parser also allows effective and *unambiguous* phrases extraction by using the internal parsing structure of `synt`, a so called *chart*. The chart is an acyclic multigraph which is built up during the analysis stage and contains all resulting trees. We have employed this technique in extracting shallow verb valencies as described further in this paper. Detailed description of the general extraction algorithm can be found in [6].

3 Verbalex Valency Lexicon

The Verbalex valency lexicon for Czech has been continuously developed since 2004. Currently it contains over 21,000 verb frames for more than 10,000 Czech verbs. It uses the notion of two-level annotation for the so called *complex valency frames* [7]: the first level provides shallow syntactic valencies (e. g. grammatical cases) whereas the second level contains deep semantic annotation using the *semantic roles*. Moreover, each valency frame contains a synonymic set mapped to the Czech WordNet [8]. In this paper we are concerned only with the first (syntactic) level. An example valency frame for the verb *skákat (to jump)* looks as follows:

3.1 The BRIEF Format

The Verbalex valency lexicon is stored primarily as XML documents, however the `BRIEF` format [9] can also be used for describing shallow syntactic valencies. The phrases extraction functionality of the `synt` parser mentioned above allows us to extract all possible valencies of the verb in this `BRIEF` format. Thanks to this, we can compare the output of the `synt` parser with the shallow valencies as recorded in the Verbalex lexicon. Sample output of the `synt` valency extraction in the `BRIEF` format is provided in the following example:

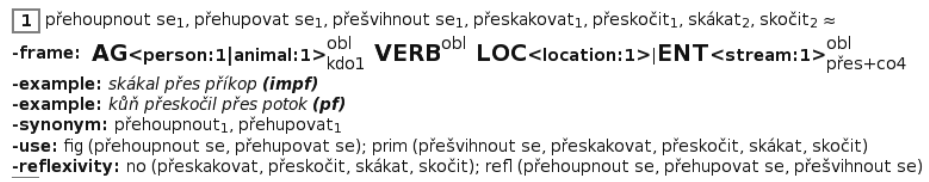

Fig. 1. Example valency frame for the Czech verb *skákat (to jump)*: it shows a nominative valency with semantic role of a `person` or an `animal` and an accusative valency with the preposition *přes (over)* and a semantic role of a `location` or `stream`.

```
; extracted from sentence: Nenadálou finanční krizi musela podnikatelka
řešit jiným způsobem .
řešit <v>hTc4a-hTc7
```

*(The businessman had to **solve** the sudden financial crisis in another way.)*

An accusative and instrumental valency has been found.

```
; extracted from sentence: Hlavní pomoc ale nacházela v dalších
obchodních aktivitách .
nacházet <v>hTc4-hTc6r{v}
```

*(However she **found** the main help in further business activities.)*

An accusative and ablativ valency with the preposition *v (in)* has been found.

```
; extracted from sentence: U výpočetní techniky se pohybuje v rozmezí od
8000 Kč do 16000 Kč .
pohybovat <v>hTc2r{u}-hTc6{v}
```

*(For information technology [it] **ranges** between 8000 Kč and 16000 Kč.)*

A genitive valency with the preposition *u (for)* and an instrumental valency with the preposition *v (in)* has been found.

4 Extraction of Shallow Valencies

As outlined above, we have extended the extraction functionality of the `synt` parser in a way that enables us to obtain various syntactic structures for the given corpus sentences. From these structures we construct simple shallow valency frames for every verb that we then compare to the valency frames available in the Verbalex lexicon. The exact extraction procedure is as follows:

1. identify clauses in the input sentence and process each of them separately,
2. in each clause, identify all prepositional and noun phrases, infinitives and selected conjunctions recorded in Verbalex (*až, že, jestli, zda, ať, aby, jak*),
3. construct a valency frame in the `BRIEF` format using the above information together with available morphological annotation,
4. for all automatically extracted valencies of a particular verb, check whether they are available in Verbalex.

In the way described above we are able to find incomplete valency frames, suggest valencies for missing verbs in Verbalex, or offer examples with most complete valency frames. The number of complete valency frames enables us

Table 1. Coverage of the Verbalex valency lexicon on the annotated DESAM corpus

indicator	covered	total	%
verb coverage	2,957	3,685	**80.24**
valency coverage	5,348	9,430	56.71
valency coverage with consideration of error analysis	5,348	6,397 [1]	**83.60**

Table 2. Error analysis of missing valencies.

indicator	number of missing valencies	%
noun phrases (*case* only valencies)	499	11.00
prepositional phrases (*preposition+case* valencies)	3,142	76.97
other (*subordinated clauses, infinitives* etc.)	491	12.03
total	4,082	100

also to determine the minimal coverage of the Verbalex lexicon with regard to the data in the DESAM corpus. The related results are shown below – note that for the purpose of this measurements, several relaxations have been performed. We matched the valency frame only against those valencies that can potentially be found by the `synt` extraction (as listed in Point 2 of the above enumeration). We also ignored the animacy denoted in the `BRIEF` format since there is no way how `synt` could obtain this information (besides the animacy of Czech masculines), i. e. the `hP` and `hT` tags have been considered to be equal and finally we also didn't differentiate between various meanings of a single verb denoted in Verbalex since there is no way how to do this on the syntactic level.

An automatic extraction of valencies performed in the way described above has one obvious drawback: with the available syntactic information we are generally not able decide whether a noun or prepositional phrase associated with a verb is an obligatory valency (argument) or a non-obligatory adjunct (modifier) usually expressing time, place or manner (in most cases with a preposition). Therefore we performed a manual error analysis of a random sample of 200 potential prepositional valencies which have not been found (representing over 75 % from all missing valencies) in the Verbalex lexicon. It revealed that a vast majority (193, i. e. 96.5 %) of those valencies were actually adjuncts (deliberately not listed in the Verbalex lexicon).

Thus our results provided in Table 1 below consist of three different measurements. First, we show the coverage of the lexicon on a verb-only level (i. e. whether there is an entry for a verb in the lexicon). Second, we show the raw results of valency matching (performed on the verbs available in the lexicon). In the end we give an estimation of the valency coverage by extrapolating our error analysis results on missing prepositional valencies to the whole lexicon. Based on this measurement, we estimate the minimum coverage of the Verbalex lexicon on valency level to be 83.60 %.

[1] The extrapolated value from the error analysis has been computed as: $9430 - 0.965 \cdot 3142$ *(number of all potential valencies − relative frequency of adjuncts in the sample · number of all missing prepositional valencies).*

5 Conclusion

In this paper we described the involvement of the Czech parser `synt` in developing and evaluating of the Verbalex valency lexicon. We consider the demonstrated method as well as the underlying technique (extraction of phrases) to be universal and easily applicable to similar tasks in the future. It should be also mentioned that the relation between `synt` and Verbalex is symbiotic in many aspects: we can not only improve Verbalex by `synt`, but also enhance the parser with the help of Verbalex, as it has been proposed in [10]. Our results support the evidence that verb valencies play an integral role in Czech syntax and should be further investigated.

Acknowledgements This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536 and in the National Research Programme II project 2C06009.

References

1. Horák, A., Pala, K.: Building a large lexicon of complex valency frames. In: Proceedings of the FRAME 2007: Building Frame Semantics Resources for Scandinavian and Baltic Languages, Lund University, Sweden, Tartu, Estonia (2007) 31–38.
2. Horák, A., Holan, T., Kadlec, V., Kovář, V.: Dependency and phrasal parsers of the Czech language: A comparison. In: Proceedings of the 10th International Conference on Text, Speech and Dialogue, Pilsen, Czech Republic, Springer Verlag (2007) 76–84.
3. Pala, K., Rychlý, P., Smrž, P.: DESAM – Annotated Corpus for Czech. In: Proceedings of SOFSEM '97, Springer-Verlag (1997) 523–530.
4. Kadlec, V., Horák, A.: New meta-grammar constructs in czech language parser synt. In: Lecture Notes in Computer Science, Springer Berlin / Heidelberg (2005).
5. Kadlec, V.: Syntactic analysis of natural languages based on context-free grammar backbone. Ph.D. thesis, Faculty of Informatics, Masaryk University, Brno (2007).
6. Jakubíček, M., Horák, A., Kovář, V.: Mining phrases from syntactic analysis. In: Proceedings of the 12th International Conference on Text, Speech and Dialogue, Pilsen, Czech Republic, Springer Verlag (2009) 124–130.
7. Pala, K., Horák, A.: Can complex valency frames be universal? In: RASLAN 2008, Brno, Masarykova Univerzita (2008) 41–48.
8. Pala, K., Smrž, P.: Building Czech wordnet. In: Romanian Journal of Information Science and Technology, Romanian Academy (2004) 79–88.
9. Pala, K., Ševeček, P.: The valencies of Czech words. In: Sborník prací FFBU, Brno, Masarykova univerzita (1997) 41–54.
10. Hlaváčková, D., Horák, A., Kadlec, V.: Exploitation of the Verbalex verb valency lexicon in the syntactic analysis of Czech. In: Proceedings of Text, Speech and Dialogue 2006, Brno, Springer-Verlag (2006) 85–92.