Exploitation of the VerbaLex Verb Valency Lexicon in the Syntactic Analysis of Czech

Dana Hlaváčková, Aleš Horák, and Vladimír Kadlec

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

Abstract. This paper presents an exploitation of the lexicon of verb valencies for the Czech language named VerbaLex. The VerbaLex lexicon format, called *complex valency frames*, comprehends all the information found in three independent electronic dictionaries of verb valency frames and it is intensively linked to the Czech WordNet semantic network.
The NLP laboratory at FI MU Brno develops a deep syntactic analyzer of Czech sentences, the parsing system `synt`. The system is based on an efficient and fast head-driven chart parsing algorithm. We present the latest results of using the information contained in the VerbaLex lexicon as one of the language specific features used in the tree ranking algorithm for the Best Analysis Selection algorithm, which is a crucial part of the syntactic analyser of free word order languages.

1 Introduction

The ambiguity level in the syntactic analysis of free word order languages suffers from the exponential explosion of the number of resulting derivation trees. The main reasons for this combinatorial grow arise on several levels of the sentence building process (prepositional attachment, verb argument resolution, non-projectivity, ellipsis, anaphoric relations, etc.). A traditional solution for these problems is presented by probabilistic parsing techniques aiming at finding the most probable parse of a given input sentence. This methodology is usually based on the relative frequencies of occurrences of the possible relations in a representative corpus. "Best" trees are judged by a probabilistic figure of merit. Our experiments show, that in the case of really free word order languages (like Czech) the probabilistic measures are not able to cover the complexity of the sentence syntax. That is why we need to exploit the knowledge of the language specific features as described in [1].

The basic sentence frame is driven by the lexical characteristics of its predicative construction based on the set of possible verb valencies of the particular verb (see e.g. [2]). We have implemented the technique of discovering the possible verb valencies from the resulting ambiguous packed shared forest (stored in the parsing chart). This enables us to work with verb valencies in two directions: a) using the VerbaLex valency lexicon to prune impossible combination regarding

the particular verb, and b) automatically process large corpora for discovering possible verb valencies that are missing in the lexicon. These valencies are then offered to the linguistic expert for addition to VerbaLex. Similar approach has been described in [3], in which a partial parsing outputs were used for obtaining the verb subcategorization information. Our approach includes a full parsing of Czech sentence, which increases the credibility of the verb frame information.

2 The VerbaLex Valency Lexicon

This paper presents an exploitation of the lexicon of verb valencies for the Czech language named VerbaLex [4]. VerbaLex was created in 2005 and it is based on three valuable language resources for Czech, three independent electronic dictionaries of verb valency frames.

The first resource, Czech WordNet valency frames dictionary, was created during the Balkanet project and contains semantic roles and links to the Czech WordNet semantic network. The other resource, VALLEX 1.0 [5], is a lexicon based on the formalism of the Functional Generative Description (FGD) and was developed during the Prague Dependency Treebank (PDT) project. The third source of information for VerbaLex is the syntactic lexicon of verb valencies denoted as BRIEF, which originated at FI MU Brno in 1996 [6].

The resulting lexicon, VerbaLex, comprehends all the information found in these resources plus additional relevant information such as verb aspect, verb synonymy, types of use and semantic verb classes based on the VerbNet project [7]. The information in VerbaLex is organized in the form of *complex valency frames* (CVF). All the valency information in VerbaLex is specified regarding the particular verb senses, not only the verb lemmata, as it was found in some of the sources. The current work on the lexicon data aims at enlarging the lexicon to the size of about 16.000 Czech verbs. The VerbaLex lexicon displays syntactic dependencies of sentence constituents, their semantic roles and links to the corresponding Czech WordNet classes. An example of such verb frame is presented in the Figure 1.[1]

The complex valency frame in VerbaLex is designed as a sequence of elements which form a "pattern"[2] for obligatory sentence constituents that depend on the verb. There are two types of information displayed in CVF. The constituent elements of valency frames cover both syntactic level and lexical semantic level (represented by two-level semantic roles). The default verb position 'VERB' as the centre of the sentence is marked on the syntactic level. The pattern of sentence constituents are situated in left and right positions in accordance with the complementarity needed by the verb. The constituent elements of frame entries are entered as pure pronominal terms, e.g. `kdo` (who), `co` (what), or

[1] This is a slightly enhanced version of CVF that splits the attribute values to *verb attributes* and *frame attributes*.

[2] A list of necessary grammatical features such as the grammatical case or the the preposition.

Princeton WordNet: dress:2, clothe:1, enclothe:1, garb:1, raiment:1, tog:1, garment:1, habilitate:2, fit out:2, apparel:1
definition: provide with clothes or put clothes on
VerbaLex Synset: obléci:1$_{pf}$, oblékat:1$_{impf}$, obléknout:1$_{pf}$, ustrojit:1$_{pf}$, strojit:1$_{impf}$
=def: provide with clothes or put clothes on
=canbepassive: yes
=meaning: I
=class: dress-41.1.1
Complex valency frames:

1. obléci:1, oblékat:1, obléknout:1
   -frame: AG<person:1>$^{obl}_{who1}$ VERB
   PAT<person:1>$^{obl}_{to\_whom3}$ ART<garment:1>$^{obl}_{what4}$
   -synonym: ustrojit:1, strojit:1
   -example: *maminka oblékla dítěti kabát / the mother put a coat on her child*
   -attr: use: prim, reflexivity=obj_dat, mustbeimperative=no
2. obléci:1, oblékat:1, obléknout:1, ustrojit:1, strojit:1
   -frame: AG<person:1>$^{obl}_{who1}$ VERB
   PAT<person:1>$^{obl}_{whom4}$ ART<garment:1>$^{obl}_{in+sth2}$
   -synonym:
   -example: *maminka oblékla dítě do kabátu / the mother dressed her child in a coat*
   -attr: use: prim, reflexivity=obj_ak, mustbeimperative=no

Fig. 1. An example of a VerbaLex verb frame

prepositional phrase pattern (with the lemma of the preposition) followed by the number of the required grammatical case of the phrase.

opustit:4/*leave office*:1 (give up or retire from a position)
frame: AG<person:1>$^{obl}_{who1}$ VERB ACT<job:1>$^{obl}_{what4}$
example: opustil zaměstnání / he left his job

This way of notation allows to differentiate an animate or inanimate subject or object position. The types of verbal complementation are precisely distinguished in the verb frame notation.

If a verb requires a completion with adjective or adverb, this fact is written as the adjectival or adverbial lemma and part of speech tag from WordNet semantic network – `[a]` or `[b]`.

cítit se:1/*feel*:5 (have a feeling or perception about oneself in reaction to someone's behavior or attitude)
frame: AG<person:1>$^{obl}_{who1}$ VERB ATTR<[a]>$^{obl}_{which1}$
example: cítil se bezvýznamný / he felt insignificant

cítit se:2/*feel*:4 (seem with respect to a given sensation given)
frame: AG<person:1>$^{\text{obl}}_{\text{who1}}$ VERB MAN<[b]>$^{\text{opt}}_{\text{how}}$
example: cítil se špatně / he felt badly

A verb valency with an infinitive construction is marked by abbreviation 'inf' and link to the verbal literal from Princeton WordNet. A subordinate clause complementation is specified by the lemma of the subordinating conjunction.

začít:1/*begin*:1 (take the first step or steps in carrying out an action)
frame: AG<person:1>$^{\text{obl}}_{\text{who1}}$ VERB ACT<[v]>$^{\text{obl}}_{\text{inf}}$
example: začal stavět dům / he began to build a house

popřít:1/*disclaim*:1 (renounce a legal claim or title to)
frame: AG<person:1>$^{\text{obl}}_{\text{who1}}$ VERB COM<statement:1>$^{\text{obl}}_{\text{that}}$
example: popřel, že ho zná/he disclaimed that he knows him

The type of valency relation can be obligatory 'obl' (must be present) or optional 'opt'. With this notation format it is possible to generate two (or more) frames from one basic frame.

The basic frame

děsit:1/*frighten*:1 (cause fear in)
frame: AG<person:1>$^{\text{obl}}_{\text{who1}}$ VERB PAT<person:1>$^{\text{obl}}_{\text{whom4}}$
ACT<act:2>$^{\text{opt}}_{\text{with\_what7}}$
example: děsil ho hrozbami / he frightened him with threats

contains the potential frame:

frame: AG<person:1>$^{\text{obl}}_{\text{who1}}$ VERB PAT<person:1>$^{\text{obl}}_{\text{whom4}}$
example: děsil ho / he frightened him

Other details of the complex valency frame notation (e.g. the way of selection of the two-level semantic roles that link the constituents to the wordnet hyperohyponymical hierarchy) are described in [4].

3 The Syntactic Analyzer `synt`

The NLP laboratory at FI MU Brno develops a deep syntactic analyzer of Czech sentences, the parsing system `synt` [8]. The system uses the meta-grammar formalism, which enables to define the grammar with a maintainable number of meta-rules. These meta-rules are produced manually by linguists. The rules are then translated into context-free rules supplemented with additional contextual constraints and semantic actions. Efficient and fast head-driven chart parsing algorithm is used for the context-free parsing. The result of the context-free parsing process – a chart – is stored in the form of a packed shared forest. To apply the constraints and to compute the semantic actions, we build a new *forest of values* instead of pruning the original chart. We use this multi-pass approach,

because all described functions are implemented as plug-ins that can be modified as needed or even substituted with other implementations. For example, we compared four different parsing algorithms which use identical internal data structures. The parsing system is aimed at analyzing the sentence not only at the surface level, but it also covers the logical analysis of the sentence by means of the Transparent Intensional Logic (TIL) [9].

4 The Verb Frames and Syntactic Analysis

In the case of a syntactic analysis of a really free word order language as the Czech language is, we need to exploit the language specific features for obtaining the correct ordering of the resulting syntactical analyzes. So far the most advantageous approach is the one based upon valencies of the verb phrase – a crucial concept in traditional linguistics.

The part of the system dedicated to exploitation of information obtained from a list of verb frames is necessary for solving the prepositional attachment problem in particular. During the analysis of noun groups and prepositional noun groups in the role of verb valencies in a given input sentence one needs to be able to distinguish free adjuncts or modifiers from obligatory valencies. The wordnet classes together with the surface features in complex valency frames are directly used for setting up a set of heuristic rules that determine whether a noun group found in the sentence serves here as a free adjunct or not. The heuristics are based on the lexico-semantic constraints derived from the VerbaLex links to the EuroWordNet hypero-hyponymical hierarchy.

4.1 Automatic Extraction of Verb Frames from the Packed Shared Forest

The verb frame extraction (VFE) process in the `synt` system is controlled by the metagrammar semantic actions. As we have described in the Section 3, we build a forest of values to represent a result of the application of contextual constraints. The VFE actions are then executed on a different level (see [8]) than the "usual" actions, which allows us to apply VFE actions on the whole forest of values.

First of all, we find all noun groups covered by the particular context-free rule. Then compatible groups[3] are processed by the VFE action. Notice, that this step suffers from a possible exponential time complexity because we work with the derivation trees and not with the packed forest. On the other hand our experiments show (see the Table 1) that in the average case this is not a problem.

If the analyzed verb has a corresponding entry in VerbaLex, we try to match the extracted frame with frames in the lexicon. When checking the valencies with VerbaLex, the dependence on the surface order is discharged. Before the

[3] Compatible in the term of derivation, i.e. groups within the same derivation tree.

system confronts the actual verb valencies from the input sentence with the list of valency frames found in the lexicon, all the valency expressions are reordered. By using the standard ordering of participants, the valency frames can be handled as sets independent on the current position of verb arguments. However, since VerbaLex contains an information about the *usual* verb position within the frame, we promote the standard ordering with increasing or decreasing the respective derivation tree probability.

We have measured the results of the first version of the automatic verb frame extraction on 4117 sentences from the Czech corpus DESAM [10]. We have selected sentences which are analysed on the rule level 0, i.e. sentences, which do not contain analytically difficult phenomena like non-projectivity or adjective noun phrase. Even on those sentences the number of possible valency frames can be quite high (see the Table 1). However, if we work with intersections of those possible valency frames, we can get a useful reduction of the number of resulting derivation trees – see the examples described in the next Section.

Table 1. The results of verb frame extraction from the corpus DESAM.

Number of sentences:	
count	**4117**
Number of words in sentence:	
minimum	2.0
maximum	68.0
average	16.8
median	**15.0**
Number of discovered valency frames:	
minimum	0
maximum	37080
average	380
median	**11**
Elapsed time:	
minimum	0.00 s
maximum	274.98 s
average	6.86 s
median	**0.07 s**

4.2 Examples

The projection of the extracted valency frames to the corresponding VerbaLex entry can be used as effective pruning tool for decreasing the number of successful derivation trees. As an example of such pruning, we can have a look at the sentence

Pokud *uchazeči kurs rekvalifikace úspěšně absolvují*, budou mít jistě uplatnění v zaměstnání.
If *the candidates successfully complete the retraining course*, they will certainly assert themselves in their job.

The valency frame for the verb 'absolvovat' from VerbaLex:

absolvovat:1/*complete*:1 (come or bring to a finish or an end)
AG<person:1>$^{\text{obl}}_{\text{who1}}$ VERB KNOW<course:1>$^{\text{obl}}_{\text{what4}}$

There are 132 trees for that sentence in the parsing system `synt`. Due to the free word order the sequence of sentence parts is

subject (uchazeči/candidates) – *object* (kurs/course)
– *verb* (absolvují/complete).

According to the valency frame the subject is a noun in nominative and the object is a noun in accusative. It is evident, that those elements cannot form a nominal phrase. This constriction reduces the number of trees to 24.

Another example is displayed in the following sentence:

Havel se radil s představiteli justice a vnitra o posílení práva.
Havel consulted with representatives of judiciary and home office on the consolidation of the legal system.

The valency frame for the verb 'radit se' from VerbaLex is:

radit se:1/*consult*:1 (get or ask advice from)
AG<person:1>$^{\text{obl}}_{\text{who1}}$ VERB SOC<person:1>$^{\text{opt}}_{\text{with\_whom7}}$
ENT|ABS <entity:1,abstraction:1>$^{\text{opt}}_{\text{about\_what6}}$

The number of `synt` trees for this sentence is 2672. The part of sentence with the preposition 's' (with) and a noun in instrumental and the part of sentence with preposition 'o' (on) and a noun in locative are necessarily prepositional nominal phrases. The application of such limits in `synt` allows a significant reduction of the number of trees to 18.

5 Conclusions

We have presented the results of exploitation of automatic verb frame extraction for Czech as a language specific feature used for pruning the packed shared forest of results of syntactic analysis with the `synt` parser. A necessary tool for this, the VerbaLex lexicon of valency frames that is being built at FI MU Brno, is also described.

The preliminary results of the exploitation of VerbaLex in the syntactic analysis of Czech are very promising and the precision of the analysis grows significantly. We believe that with enlarging the lexicon to a representative number of Czech verbs the `synt` system will be able to detect the correct derivation tree in many cases which were unsolvable so far.

Acknowledgments

This work has been partly supported by Czech Science Foundation under the project 201/05/2781 and by Grant Agency of the Academy of Sciences of CR under the project 1ET400300414.

References

1. Horák, A., Smrž, P.: Best analysis selection in inflectional languages. In: Proceedings of the 19th international conference on Computational linguistics, Taipei, Taiwan, Association for Computational Linguistics (2002) 363–368
2. Trueswell, J., Kim, A.: How to prune a garden-path by nipping it in the bud: Fast-priming of verb argument structures. Journal of Memory and Language (39) (1998) 102–123
3. Gamallo, P., Agustini, A., Lopes, G.P.: Learning subcategorisation information to model a grammar with co-restrictions. Traitement Automatique de la Langue **44**(1) (2003) 93–117
4. Hlaváčková, D., Horák, A.: Verbalex – new comprehensive lexicon of verb valencies for czech. In: Proceedings of the Slovko Conference, Bratislava, Slovakia (2005)
5. Žabokrtský, Z., Lopatková, M.: Valency Frames of Czech Verbs in VALLEX 1.0. In Meyers, A., ed.: HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation. (2004) 70–77
6. Pala, K., Sevecek, P.: Valence českých sloves (Valencies of Czech Verbs). In: Proceedings of Works of Philosophical Faculty at the University of Brno, Brno, Masaryk University (1997) 41–54
7. Dang, H.T., Kipper, K., Palmer, M., Rosenzweig, J.: Investigating regular sense extensions based on intersective levin classes. In: Proceedings of Coling-ACL98, Montreal CA (August 11-17, 1998) www.cis.upenn.edu/ mpalmer/.
8. Horák, A., Kadlec, V.: New meta-grammar constructs in czech language parser synt. In: Proceedings of Text, Speech and Dialogue 2005, Karlovy Vary, Czech Republic, Springer-Verlag (2005) 85–92
9. Horák, A.: The Normal Translation Algorithm in Transparent Intensional Logic for Czech. PhD thesis, Faculty of Informatics, Masaryk University, Brno (2002)
10. Pala, K., Rychlý, P., Smrž, P.: DESAM — annotated corpus for Czech. In: Proceedings of SOFSEM'97, Springer-Verlag (1997) 523–530 Lecture Notes in Computer Science 1338.