# Efficient Sentence Parsing with Language Specific Features: A Case Study of Czech

**Aleš Horák and Pavel Smrž**
Faculty of Informatics, Masaryk University Brno
Botanická 68a, 602 00 Brno, Czech Republic
E-mail: {hales,smrz}@fi.muni.cz

**Abstract**

This paper presents two optimalizations of standard parsing techniques applied to Czech as a representative of a free word order language with rich morphology. Our approach is based on features that take advantage of particular knowledge of language specific properties. The first method involves a special type of merging possible due to the agreement fulfillment. Another use of language specific features lies in the exploitation of verb valency frames for probabilistic ordering of parsing output.

## 1 Introduction

Parsing of natural language texts finds its use in many areas of computational linguistics such as machine translation, speech recognition or text summarization. Results of analysis of English have already reached a satisfactory level to be applied in real-world applications.

However, the same parsing techniques working with a morphologically rich free word order language like Czech suffer from serious problems caused by a high ambiguity of the underlying grammar. Syntactical analysis of Czech usually yields enormously high number of possible parsing trees, which together with inevitable feature structure unification require an unacceptable computational resources. We discuss two methods to overcome these difficulties.

The next section describes the use of a special feature structure designed to effectively merge relevant syntactic features, which prevents from a combinatorical expansion of possible unification outputs. In section 3, we present a technique that supplements standard stochastic methods with verb valency concept to be able to capture the free word order.

## 2 Language Specific Feature Merging

In a general case the number of unicification actions grows exponentially with the level of parsing ambiguities in the resulting chart. One approach that may (with some kinds of input structures) reduce the number of actions lies in utilization of the interleaved pruning actions. However, many natural language phenomena defy to solving a significant portion of the ambiguities on the local context involved in interleaved pruning.

Instead of that, an efficient version of non-interleaved pruning (factored extraction [5]) can be used to remove some overabundant combinations of unified values and by this way decrease the number of unifications after the chart has been built. This technique is limited to combinations of evincibly identical constraints.

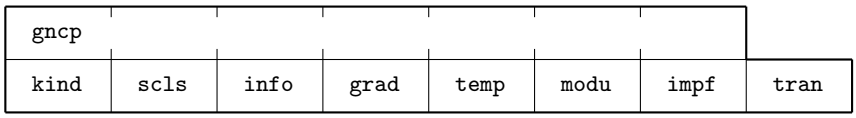


Figure 1: Structure used for Czech language specific feature merging

Another technique, used in our system, makes the best of the knowledge of the language specific features used in agreement fulfillment tests. It exploits the fact that in local unifications the core feature structures of the right hand side constituents can be merged together in the left hand side feature structure describing all possible variants of the grammatical features of the result. For the Czech language we use a structure described in the Figure 1.

The feature merging mostly comprises the `gncp` part of the structure which describes:

- for a *noun phrase* all its possible grammatical cases (in Czech at most 7 cases — nominative, genitive, dative, accusative, vocative, locative and instrumental), numbers (singular and/or plural) and genders (out of 4 possible — masculine animative and inanimative, feminine and neutral).
- for a *verb phrase* all its possible numbers, genders and persons (up to 3).

The `gncp` part represents a boolean array indicating the presence or absence of the specified grammatical feature in the given noun/verb phrase. This structure can thus express up to 56 ($7 \cdot 2 \cdot 4$) combinations of feature values.

The other fields in the feature structure represent various grammatical features of the constituents.

# 3 Verb Valencies as a Figure of Merit

Ambiguity on all levels of representation is an inherent property of natural languages and it also forms a central problem of natural language parsing. A consequence of the natural language ambiguity is a high number of possible outputs of a parser that are usually represented by labeled trees. The average number of parsing trees per input sentence strongly depends on the background grammar and thence on the language. There are natural language grammars producing at most hundreds or thousands of parsing trees but also highly ambiguous grammar systems producing enormous number of results. For example, a grammar extracted from the Penn Treebank and tested on a set of sentences randomly generated from a probabilistic version of the grammar has on average $7.2 \times 10^{27}$ parses per sentence according to Moore's work (IWPT'2000). Such a mammoth extent of result is also no exception in parsing of Czech [8] (see Fig. 2) due to free word order and rich morphology of word forms which grammatical case cannot often be unambiguously determined.

A traditional solution for these problems is presented by probabilistic parsing techniques [2] aiming at finding the most probably parse of a given input sentence. This methodology is usually based on the relative frequencies of occurrence of the possible relations in a representative corpus. "Best" trees are judged by a probabilistic figure of merit.

A key question is then what the good candidates for FOMs are. The use of probabilistic context-free grammars (PCFGs) involves simple CF rule probabilities to form a FOM [4, 1]. Caraballo and Charniak [3] present and evaluate different figures of merit in the context of best-first chart parsing. They recommend boundary trigram estimate that has achieved the best performance on two testing

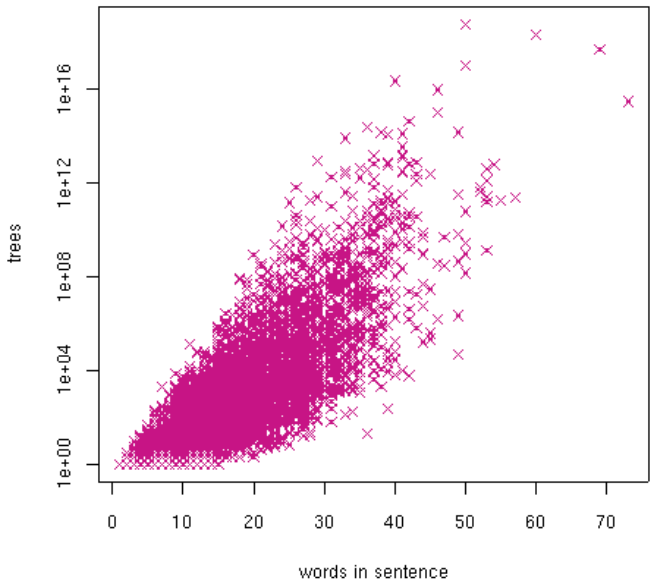


Figure 2: The level of ambiguity grows exponentially to the number of words in a sentence.

grammars. This technique, as well as stochastic POS tagging based on $n$-gram statistics, achieves satisfactory results for analytical languages (like English). However, in case of free constituent order languages, current studies suggest that these simple stochastic techniques considerably suffer from the data sparseness problem and require a formidable amount of training data.

In order to cope with these difficulties in Slavonic languages (viz. Czech), we propose to exploit the language specific features. Preliminary results indicate that the most advantageous approach is the one based upon valencies of the verb phrase — a conception often discussed in traditional linguistics.

Certainly, we need to discharge the dependence on the surface order. The first step lies in computation of $n$-grams based on the lexical heads. On one hand it follows the principle of free word ordering and on the other hand it reduces the number of possible training schemata, which may be crucial to the usability. After that, we employ a mechanism that relaxes the requirement of $n$-gram to form a tuple with firm order. In our definition, the $n$-gram is counted as an unordered collection of items (internally, the set is represented as a list with predefined ordering of its elements, so that two sets that contain the same items are equal abstractedly from their order).

The part of the system dedicated to exploitation of information obtained from a list of verb valencies [6] 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. We are testing a set of heuristic rules that determine whether a found noun group typically serves as a free adjunct. The heuristics are based on the lexico-semantic constraints [7].

All these language specific optimalizations of standard methods have a substantial impact on the overall performance. The Table 1 summarizes the precision estimates counted on real corpus data.

| | percentage |
|---|---|
| precision on sentences of 1-10 words | 86.9 % |
| precision on sentences of 11-20 words | 78.2 % |
| precision on sentences of more than 20 words | 63.1 % |
| overall precision | **79.3 %** |
| number of sentences with mistakes in input | 8.0 % |

Table 1: Precision estimate as per sentence length

These measurements presented here may underestimate the actual benefits of this approach due to the estimated 8% of mistakes in input corpus.

# 4 Conclusions

Both techniques described in the article show that the utilization of language specific features form a substantial asset to standard parsing approaches. To the best of our knowledge, the presented results represent the first-rate values of precision and time requirements that has been achieved for Czech so far. The advantage of these methods lies in their effortless integrability with lexical associations, which should provide in-depth domain knowledge represented by a specialized corpus.

# References


[1] Robert J. Bobrow. Statistical agenda parsing. In *Proceedings of the February 1991 DARPA Speech and Natural Language Workshop*, pages 222–224. San Mateo: Morgan Kaufmann.

[2] Harry Bunt and Anton Nijholt, editors. *Advances in probabilistic and other parsing technologies.* Kluwer Academic Publishers, 2000.

[3] Sharon Caraballo and Eugene Charniak. New figures of merit for best-first probabilistic chart parsing. *Computational Linguistics*, 24(2):275–298, 1998.

[4] M. Chitrao and R. Grishman. Statistical parsing of messages. In *Proceedings of the Speech and Natural Language Workshop*, pages 263–266, Hidden Valley, PA, 1990.

[5] J. T. Maxwell III and R. M. Kaplan. The interface between phrasal and functional constraints. In M. Rosner, C. J. Rupp, and R. Johnson, editors, *Proceedings of the Workshop on Constraint Propagation, Linguistic Description, and Computation*, pages 105–120. Instituto Dalle Molle IDSIA, Lugano, 1991. Also in Computational Linguistics, Vol. 19, No. 4, 571–590, 1994.

[6] Karel Pala and Pavel Ševeček. Valencies of czech verbs. In *Proceedings of Works of Philosophical Faculty at the University of Brno*, pages 41–54. Brno, 1997. (in Czech).

[7] Pavel Smrž and Aleš Horák. Implementation of efficient and portable parser for czech. In *Text, Speech and Dialogue: Proceedings of the Second International Workshop TSD'1999*, Pilsen, Czech Republic, 1999. Springer Verlag, Lecture Notes in Computer Science, Volume 1692.

[8] Pavel Smrž and Aleš Horák. Large scale parsing of czech. In *Proceedings of Efficiency in Large-Scale Parsing Systems Workshop, COLING'2000*, pages 43–50, Saarbrucken: Universitaet des Saarlandes, 2000.