V. Kovář, V. Kadlec, A. Horák

GRAMMAR DEVELOPMENT FOR CZECH SYNTACTIC PARSER WITH CORPUS-BASED TECHNIQUES

1. Introduction

The main tasks of the automatic natural language (NL) processing are based on a correct syntactic analysis of a NL sentence. While the quality of parsing of analytical languages (English, German, ...) has already achieved nearly satisfactorily results[1], the analysis of free word order languages still generates many problems either in the form of huge number of rules or output trees or offers lower precision or coverage on corpus texts[2].

In the NLP laboratory at Masaryk University in Brno Czech Republic, syntax parsing of Czech language as a representative of really free word order language forms one of the main stream tasks since the establishment of the laboratory. The most advanced system out of several implemented analysers is a parser called **synt**[3]. Recently, **synt** features a developed meta-grammar of Czech and a fast parsing mechanism which offers a coverage of

[1] *Baumann S., Brinckmann C., Hansen-Schirra S. et al.* The muli project: Annotation and analysis of information structure in German and English // Proceedings of the LREC–2004 Conference. Lisboa, Portugal, 2004; *Radford A.* Minimalist Syntax. Cambridge University Press, Chicago, 1993.

[2] *Horák A., Smrž P.* Best analysis selection in inflectional languages // Proceedings of the 19th International Conference on Computational Linguistics. Taipei, Taiwan, Association for Computational Linguistics, 2002. P. 363–368; *Jaeger T., Gerassimova V.* Bulgarian word order and the role of the direct object clitic in LFG // *Butt M., King T.* (eds.) Proceedings of the LFG02 Conference. Stanford, CSLI Publications, 2002; *Hoffman B.* The Computational Analysis of the Syntax and Interpretation of Free Word Order in Turkish. PhD thesis, University of Pennsylvania, Philadelphia, 1995.

[3] *Horák A., Kadlec V.* New meta-grammar constructs in Czech language parser **synt** // Proceedings of Text, Speech and Dialogue–2005, Karlovy Vary, Czech Republic, Springer-Verlag, 2005. P. 85–92.

more than 92% of Czech sentences while keeping the analysis time on the average of 0,07s/sentence. The current development of the parser is aimed at exploring several *best analysis selection* methods that allow to identify the preferred derivation tree in the output – due to the grammar ambiguity the number of possible outputs (derivation trees) in **synt** is often very high (in some cases hundreds of thousands and more). For many reasons, this number is unacceptable and so methods aimed at reducing it (while keeping the precision of the output) must be developed.

2. The Grammar Development Process

The **synt** meta-grammar (denoted as grammar form G1) is carefully developed by linguistic experts – currently it contains about 300 meta-rules plus special generative constructs and selectional restrictions[1]. For the actual parsing, this grammar form is automatically expanded to one of the generated forms G2 or G3 (with about 3000 or 11000 rules), which describe the context free backbone supplemented with additional tests and actions for capturing the context dependencies.

For supporting of automatic consistency checks of the grammar development, a new tree-bank (corpus of «trees») of **synt** derivation trees is being developed. Any changes in the meta-grammar then undergoes an evaluation against this tree-bank. All phases of this process are controlled with the Grammar Development Workbench (GDW[2]) tool, which provides all necessary functions for the linguistic experts working with the system.

---

[1] *Horák A., Kadlec V.* New meta-grammar constructs in czech language parser **synt**...; *Horák A., Smrž P.* Efficient sentence parsing with language specific features:a case study of Czech // Proceedings of the IWPT–2001 (the 7th International Workshop on Parsing Technologies). Beijing, Peking University, 2001. P. 221–224.

[2] *Horák A., Kadlec V.* Platform for full-syntax grammar development using meta-grammar constructs // Proceedings of the Paclic 2006. Hubei, China, Huazhong Normal University, 2006.

GDW includes five modules – the GUIsynt graphical user interface for parsing, the TreeView and the ChartView for inspection of the analysis results, the GrammarView and the TBManager tree-bank manager.

The development of the parser goes hand in hand with the development of the meta-grammar format. For decisions about the order of implementation of proposed enhancement methods, different testing corpora and tree-banks are used for statistical estimates of the payoff of the particular method. An example of such tree corpus utilization is presented in the following Section.

3. Using Corpora to Guide the Parser Development

The particular method tested in this paper has been given a working name *beautified chart* method. Its principal idea consists in automatic transformation of the resulting derivation chart (denoted also as *packed shared forest*) to new chart structure that hides differences in the parsing that have low (or only «technical») impact to the linguistic interpretation of the results.

The aim of the corpus testing of the method is to find out how much beautified chart method reduces the number of possible output trees. The design of the used transformation operations follows the requirement of *reducing* the chart, so the number of resulting («beautified») trees is supposed to be lower than numbers of the original full trees. The task is to find out how dramatically the number of derivation trees decreases, if there is some correlation with the sentence length etc.

In the test, we have randomly selected 5000 Czech sentences from 3 to 52 words. These sentences were extracted from different sources (web pages, books, newspapers, ...). After processing the sentences with **synt** the number of possible resulting «beautified» trees were computed, which allows to estimate the asset of the method (for this computation we have limited the number of output trees to 50 000). The new statistical quantity *Removed trees percentage* was defined as $100*(nf-nb)/nf$ where nf = number of full trees, nb = number of «beautified» trees.

4. Results and Interpretation

Table 1. Results – all sentences

	#sentences	percentage
Accepted for testing	3393	67.86%
Excluded from testing	877	17.54%
Not accepted	730	14.60%
Total sentences	5000	100.00%

Table 2. Results – per sentence length (for short, medium and long sentences)

	short sentences (maximum 12 words)		*medium-length sentences (13 – 20 words)*		*long sentences (more than 20 words)*	
	#sentences	percentage	#sentences	percentage	#sentences	percentage
Accepted for testing	1351	92,72%	1432	84,09%	610	33.15%
Excluded from testing	0	0,00%	86	5,05%	791	42.99%
Not accepted	106	7,28%	185	10,86%	439	23.86%
Total sentences	1457	100,00%	1703	100,00%	1840	100.00%

Table 3. Statistical characteristics of the *Removed trees percentage*

Number of sentences	3393	*n*
Average	21,513%	$a = \frac{1}{n} \sum_{i=1}^{n} x_i$
Maximum Minimum Median	94,309% 0,000% 8,333%	
Variance	629,649%	$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - a)$
Standard deviation	25,093%	$s = \sqrt{s^2}$

In the method testing, the sentences were divided into three groups according to the parsing results (see Table 1):

- Accepted for testing – the sentence was accepted by the parser, number of full trees was lower than 50 000 («Removed trees percentage» was computed only for these trees).
- Excluded from testing – the sentence was accepted by the parser, number of full trees was higher than 50 000.
- Not accepted – the sentence was rejected by the parser.

The parser has accepted 85% of the sample sentences, two thirds of it were acceptable for our measurements. We can see in Table 2 that the parser is (naturally) more successful on short sentences. For sentences with more than 20 words, only 33% of sentences were appropriate for our testing. But since we expect the *Removed trees percentage* to be similar for longer sentences, this lower number of long sentences in the measurements should not affect the results.

Values of the *Removed trees percentage* oscillate between 0% and 94%, the average is around 8%. High variance and standard deviation indicate that separate results can significantly differ. If only long sentences are taken into consideration, the average and the median are higher. Correlation coefficients also indicate that there is a weak positive dependency especially between number of all trees and removed trees percentage. Then, we can expect higher efficiency at sentences with more than 50 000 full derivation trees. Correlations between number of full trees and sentence length are also weak which means that there are other significant factors that influence the number of full trees, especially number of possible morphological analysis of each word.

5. Conclusions

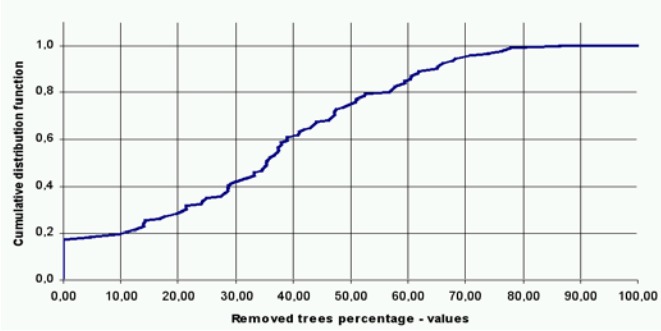
The development of quality syntax analyser is always dependent of the language resources available. In this paper, we have presented a small example of using corpus data for the grammar and parser development decisions.

The measured *Removed trees percentage* (using the application of the beautified chart method) is relatively significant (at longer sentences more than 40%):

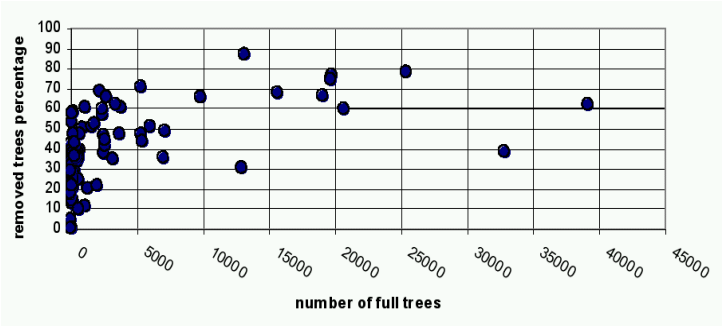
- there is a decrease of 35–43% with 50% probability;
- there is a probability of about 16% that there will be no tree reduction at all;
- there is a probability of 84% that the number of derivation trees will be reduced and 80% that the number of derivation trees will be reduced of more than 10%.

Figure 2. Graphical representation
of the *Removed trees percentage* characteristics

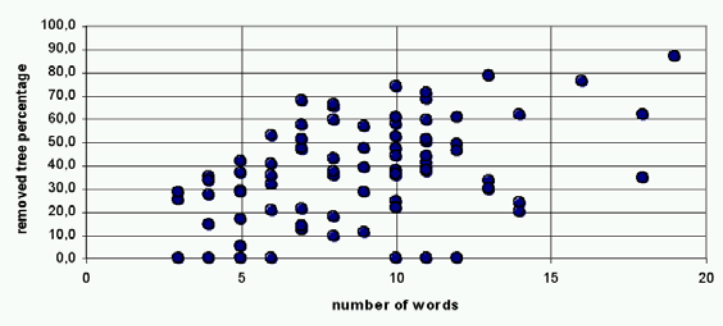
2.1. Removed trees percentage – cumulative distribution function



2.2. Removed trees percentage and sentence length



2.3. Removed trees percentage and number of full trees



Acknowledgments

This work has been partly supported by Czech Science Foundation under the project 201/05/2781, by the Academy of Sciences of CR under the project T100300414, and by the Ministry of Education of CR within the center of basic research LC05113202.