FI MU

Faculty of Informatics
Masaryk University

Corpus-based Rules for Czech Verb Discontinuous Constituents

by

Eva Žáčková
Karel Pala

FI MU Report Series FIMU-RS-99-03

 August 1999

Corpus-based Rules for Czech Verb Discontinuous Constituents*

Eva Žáčková and Karel Pala

Faculty of Informatics, Masaryk University Brno
Botanická 68a, 602 00 Brno, Czech Republic
E-mail: {glum,pala}@fi.muni.cz

August 2, 1999

Abstract

In this paper we present a method for extracting general structures of the verb groups from a tagged and fully disambiguated corpus and consecutive exploitation of these structures for the building a formal grammar in the Prolog DCG fashion. Our goal is to apply them as a rules for the analysis of the Czech verb groups in the non-disambiguated grammatically tagged Czech corpus texts. The problem of the recognition of verb discontinuous constituents in Czech is also approached and obtained statistical data are presented.

1 Introduction

For various applications in the field of natural language processing it is necessary to have rules capable to recognize and analyze reliably the verb groups in texts. The previous attempts to build the rules describing Czech verb constituents have been based rather on the introspective approach, which can hardly cover all the relevant cases occurring in the corpus texts.

*This research has been partially supported by the Czech Ministry of Education under the grant VS97028

Thus we try to arrive at a more complete algorithmic description of the Czech verb groups that could serve as a ground on which such rules can be built.

In the presented research we use corpus based techniques to find all the verb groups in DESAM corpus (Czech, general, tagged and fully disambiguated corpus containing at present about one million positions [1, 2]). The obtained results have been classified into several clusters according to their structure. Then the representative Prolog rule for each cluster has been created.

We are also touching the problem of the recognition of verb discontinuous constituents in Czech: without having it solved hardly any reasonable syntactic parser can be built. This means that it is inevitable to have a usable algorithm that would enable us to recognize such verb groups since DCG formalism in Prolog does not offer a reasonable solution of this problem. There are gapping grammars [3] that address this issue but we have not met any implementation that could yield realistic and acceptable results for Czech.

2 Extraction of the Structures of Verb Groups

The occurrences of the verb groups (from which the general structures have been extracted afterwards) were searched out in the above mentioned DESAM corpus using CQP (Corpus Query Processor [4]) queries satisfying the following conditions:

- the particular occurrence of the verb group is found exactly once,
- the whole group is found (not only a part of it),
- two independent groups are not merged together,
- the improper words (not belonging to the verb group) are not included into the group,
- groups consisting of discontinuous constituents are found.

The queries constructed take advantage of the properties of the Czech verb groups [5, 6]:

- their components are either verbs or the pronoun *se* (*si*)[1]
- a verb group cannot cross the boundary of a sentence but between its two components there can be a gap consisting of an arbitrary number of non-verb words or even a clause.

The queries were divided into several categories according to the number of "clusters" creating a searched verb group. A cluster consists of an arbitrary number of continuous components forming a verb group (see e.g. the verb group in the sentence *To **by měl** tento zákon **umožnit**.* (*It should be enabled by this law.*) consists of two clusters. Every continuous verb group consists of one cluster.).

3 Results and Statistics

From the found verb groups we extracted about 1 500 different structures which after the necessary generalizations produced about 150 skeletons of rules with the following format:

```
mít/k5e?p?n?tMmPa?  k5e?p?n?tPmCa?  sebe/k3xXn?c?
gap k5e?mFa?
```

The notation similar to that produced by morphological tagger LEMMA [7] is used. The base form of the word (can be omitted) is followed by a slash and tag which specifies required POS (part of speech) and grammatical categories. The small letters in the tag denote attributes (grammatical categories) and capital letters (or numbers) their values (e.g. `k5` denotes verb; question mark stands for any value). The string `gap` can be substituted by an arbitrary number of words which do not belong to the particular verb group. The following example shows a possible instance of the above mentioned rule (in the format: word/base-form/tag).

```
nemělo/mít/k5eNpNnStMmPaI by/by/k5eAp3nStPmCaI
se/sebe/k3xXnSc4 vedení/vést/k1gNnSc1
Telecomu/Telecom/k1gInSc2 spíše/spíše/k6xMeA
zaměřit/zaměřit/k5eAmFaP
```

(literary translation: *wouldn't the management of Telecom better focus...*)

[1]In this research we do not consider the complete verb groups consisting of the copula (mostly with *být* (*to be*)) and the noun, adjective or adverbial group. This problem will be solved in the course of the further and more complex analysis.

Verb groups found in the corpus have been also exploited for extracting some statistical data. Especially we focused on the discontinuous verb groups and the examination of gaps. Table 1 shows the representation of verb groups classified by number of its components. The second column contains the ratio of occurrences of each category in the corpus, the third column shows the percentage of the discontinuous groups of each particular category. From the table can be observed that the groups consisting of at most five components were found (we estimate that this is probably an upper limit for Czech language).

The discontinuous verb groups found in the corpus represent about 50 % of all verb groups consisting of two or more components.

# of comp.	% of all	% with gaps	example
1	61.7	–	***pršelo***
2	30.0	47.5	***se*** *brzy* ***rozhodne***
3	7.2	57.3	***chtěl jsem*** *u toho* ***být***
4	1.0	60.6	***měli bychom si*** *to* ***uvědomit***
5	0.1	66.7	***mohl by se*** *plyn* ***začít hromadit***

Table 1: Representation of the verb groups classified by the number of its components.

The frequencies of the different word types which occur in the gaps can be found in Table 2. When a sentence occurred in a gap, it was counted only once (category "sentence"), its components were not evaluated separately. The average size of a gap is two words.

The largest gap found consists of the ten components (gaps including sentences are not taken into account here): ***je*** *pověra o léčebných schopnostech tohoto preparátu v čínské medicíně hluboce* ***zakořeněna*** (*the superstition of therapeutic abilities of this stuff is widely spread in the Chinese medicine*).

4 Converting the Rules into Prolog

Before converting the skeletons of the rules created from corpus data to DCG rules the problem of parsing of discontinuous constituents in Prolog has to be faced. Since the DCG formalism does not help in this respect our solution of the problem uses a special D(iscontinuous)-predicate which enables us to collect words occurring in the gaps during parsing

category	per cent
noun	35.5
adjective	10.5
pronoun	12.5
numeral	1.8
adverb	16.7
preposition	14.7
conjunction	0.7
particle	6.9
abbreviation	0.2
sentence	0.4

Table 2: Components of gaps: percentage.

the verb group and then move them to the beginning of the rest of (not yet parsed) sentence (the idea of gapping grammars is employed here). The D-predicate can be extended by the specification of the words that can or cannot be skipped. It is also possible to use D-predicate to solve some free word order problems in Czech but its performance in this kind of analysis is not efficient enough yet.

The above mentioned skeleton of the rule (see page 3, section 3) converted into Prolog will have the following form (`gap` is the D-predicate):

```
verb_group(vg(Verb1,Verb2,Se,Verb3),Gap) -->
    word(Verb1,'mít',k5,_,_,_,tM,mP,_),
    word(Verb2,_,k5,_,_,_,tP,mC,_),
    word(Se,'sebe',k3,xX,_,_),
    gap(Gap),
    word(Verb3,_,k5,_,mF,_).
```

5 Conclusions

A method for extracting skeletons of the rules capable to recognize verb groups in Czech texts from tagged and fully disambiguated corpus has been presented. We have described practical usage of these skeletons for building a formal grammar in Prolog and we also suggested and implemented the mechanism for the analysis of the discontinuous constituents.

Further exploitation of created rules in various linguistic applications (e.g. together with the idea of verb valencies [8, 9]) is to be expected.

References

[1] Karel Pala, Pavel Rychlý and Pavel Smrž. DESAM — approaches to disambiguation. Technical Report FIMU-RS-97-09, Brno, 1997.

[2] Karel Pala, Pavel Rychlý and Pavel Smrž. DESAM — annotated corpus for Czech. In *Proceedings of SOFSEM'97*. Springer-Verlag, 1997.

[3] Veronica Dahl. More on Gapping Grammars. In *Proceedings of the International Conference on Fifth Generation Computer Systems*. Tokyo, 1984.

[4] Bruno Maxmilian Schulze and Oliver Christ. The CQP User's Manual. Universität Stuttgart, Stuttgart, 1996.

[5] Jan Petr et al. The Grammar of Czech III. Academia, Praha, 1987.

[6] Klára Osolsobě. Morphological Tagging of Composed Verb Forms in Corpus. Studia Minora Facultatis Philosophicae Universitatis Brunensis, Brno, 1999. (in print).

[7] Pavel Ševeček. *LEMMA — a Lemmatizer for Czech*. Brno, 1996. (manuscript).

[8] Karel Pala and Pavel Ševeček. Valencies of Czech Verbs. Studia Minora Facultatis Philosophicae Universitatis Brunensis, A45, 1997.

[9] Pavel Smrž and Eva Žáčková. New Tools for Disambiguation of Czech Texts. In *Proceedings of TSD'98*. Masaryk University, Brno, 1998.

Publications in the FI MU Report Series are in general accessible via WWW and anonymous FTP:

`http://www.fi.muni.cz/informatics/reports/`
`ftp  ftp.fi.muni.cz (cd pub/reports)`

Copies may be also obtained by contacting:

Faculty of Informatics
Masaryk University
Botanická 68a
602 00 Brno
Czech Republic