In: *Proceedings of CoNLL-2000 and LLL-2000*, pages 219–225, Lisbon, Portugal, 2000.

Recognition and Tagging of Compound Verb Groups in Czech

Eva Žáčková and **Luboš Popelínský** and **Miloš Nepil**
NLP Laboratory, Faculty of Informatics, Masaryk University
Botanická 68, CZ-602 00 Brno, Czech Republic
{glum, popel, nepil}@fi.muni.cz

Abstract

In Czech corpora compound verb groups are usually tagged in word-by-word manner. As a consequence, some of the morphological tags of particular components of the verb group lose their original meaning. We present a method for automatic recognition of compound verb groups in Czech. From an annotated corpus 126 definite clause grammar rules were constructed. These rules describe all compound verb groups that are frequent in Czech. Using those rules we can find compound verb groups in unannotated texts with the accuracy 93%. Tagging compound verb groups in an annotated corpus exploiting the verb rules is described.

Keywords: compound verb groups, chunking, morphosyntactic tagging, inductive logic programming

1 Compound Verb Groups

Recognition and analysis of the predicate in a sentence is fundamental for the meaning of the sentence and its further analysis. In more than half of Czech sentences the predicate contains the compound verb group. E.g. in the sentence *Mrzí mě, že jsem o té konferenci nevěděla, byla bych se jí zúčastnila.* (literary translation: *I am sorry that I did not know about the conference, I would have participated in it.*) there are three verb groups

<vg> *Mrzí* </vg> *mě, že* <vg> *jsem o té konferenci nevěděla* </vg>, <vg> *byla bych se jí zúčastnila.* </vg>

I <vg> *am sorry* </vg> *that I* <vg> *did not know* </vg> *about the conference, I* <vg> *would have participated* </vg> *in it.*

Verb groups are often split into more parts with so called gap words. In the second verb group the gap words are *o té konferenci* (*about the conference*). In annotated Czech corpora, including DESAM (Pala et al., 1997), compound verb groups are usually tagged in word-by-word manner. As a consequence, some of the morphological tags of particular components of the verb group loose their original meaning. It means that the tags are correct for a single word but they do not reflect the meaning of the words in context. In the above sentence the word *jsem* is tagged as a verb in present tense, but the whole verb group to which it belongs – *jsem nevěděla* – is in past tense. Similar situation appears in *byla bych se jí zúčastnila (I would have participated in it)* where *zúčastnila* is tagged as past tense while it is only a part of past conditional. Without finding all parts of a compound verb group and without tagging the whole group (what is necessary dependent on other parts of the compound verb group) it is impossible to continue with any kind of semantic analysis.

We consider a compound verb group to be a list of verbs and maybe the reflexive pronouns *se, si.* Such a group is obviously compound of auxiliary and full-meaning verbs, e.g. *budu se umývat* where *budu* is auxiliary verb (like *will* in English), *se* is the reflexive pronoun and *umývat* means *to wash.* As word-by-word tagging of verb groups is confusing, it is useful to find and assign a new tag to the whole group. This tag should contain information about the beginning and the end of the group and about the particular components of the verb group. It must also contain information about relevant grammatical categories that characterise the verb group as a whole. In (Žáčková and Pala, 1999), a proposal of the method for automatic finding of compound verb groups in the corpus DESAM is introduced. We describe here the improved method that results in definite clause grammar rules - called verb rules - that contain information about all components of a particular verb group and about tags. We describe also some improvements that allow us to increase the accuracy of verb group recognition.

The paper is organised as follows. Corpus DESAM is described in Section 2. Section 3 contains a description of the method for learning verb rules. Recognition of verb groups in annotated text is discussed in Section 4. Improvements of this method are introduced in Section 5. In Section 6 we briefly show how the tag for the compound verb group is constructed employing verb rules. We conclude with discussion (Section 7), the description of ongoing research (Section 8) and with a summary of relevant works (Section 9).

2 Data Source

DESAM (Pala et al., 1997), the annotated and fully disambiguated corpus of Czech newspaper texts, has been used as the source of learning data. It contains more than 1 000 000 word positions, about 130 000 different word forms, about 65 000 of them occurring more then once, and 1 665 different tags. E.g. in Tab. 1 the tag `k5eApFnStMmPaP` of the word *zúčastnila* (*participated*) means: part of speech (k) = verb (5), person (p) = feminine (F), number (n) = singular (S) and tense (t) = past (M). Lemmata and possible tags are prefixed by <l>, <t> respectively. As pointed out in (Pala et al., 1997; Popelínský et al., 1999), DESAM is not large enough. It does not contain the representative set of Czech sentences yet. In addition some words are tagged incorrectly and about 1/5 positions are untagged.

3 Learning Verb Rules

The algorithm for learning verb rules (Žáčková and Pala, 1999) takes as its input annotated sentences from corpus DESAM. The algorithm is split into three steps: finding verb chunks (i.e. finding boundaries of simple clauses in compound or in complex sentences, and elimination of gap words), generalisation and verb rule synthesis. These three steps are described below.

Mrzí	<l>mrzet
	<t>k5eAp3nStPmIaI
mě	<l>já
	<t>k3xPnSc24p1
,	
že	<l>že
	<t>k8xS
jsem	<l>být
	<t>k5eAp1nStPmIaI
o	<l>o
	<t>k7c46
té	<l>ten
	<t>k3xDgFnSc6
	<t>k3xOgXnSc6p3
konferenci	<l>konference
	<t>k1gFnSc6
nevěděla	<l>vědět
	<t>k5eNpFnStMmPaI
,	
byla	<l>být
	<t>k5eApFnStMmPaI
bych	<l>by
	<t>k5eAp1nStPmCaI
se	<l>se
	<t>k3xXnSc4
jí	<l>on
	<t>k3xPgFnSc2p3
zúčastnila	<l>zúčastnit
	<t>k5eApFnStMmPaP

Table 1: Example of the disambiguated Czech sentence

3.1 Verb Chunks

The observed properties of a verb group are the following: their components are either verbs or a reflexive pronoun *se* (*si*); the boundary of a verb group cannot be crossed by the boundary of a sentence; and between two components of the verb group there can be a gap consisting of an arbitrary number of non-verb words or even a whole sentence. In the first step, the boundaries of all sentences are found. Then each gap is replaced by tag `gap`.

The method exploits only the lemma of each word (nominative singular for nouns, adjectives, pronouns and numerals, infinitive for verbs) and its tag. We will demonstrate the whole process using the third simplex sentence of the clause in Tab. 1 (*byla bych se jí zúčastnila (I would have* *participated in it))*:

```
být/k5eApFnStMmPaI
by/k5eAp1nStPmCaI
se/k3xXnSc3
on/k3xPgUnSc4p3
zúčastnit/k5eApFnStMmPaP
```

After substitution of gaps we obtain

```
být/k5eApFnStMmPaI
by/k5eAp1nStPmCaI
si/k3xXnSc3
gap
zúčastnit/k5eApFnStMmPaP
```

3.2 Generalisation

The lemmata and the tags are now being generalised. Three generalisation operations are employed: elimination of (some of) lemmata, generalisation of grammatical categories and finding grammatical agreement constraints.

3.2.1 Elimination of lemmata

All lemmata except forms of auxiliary verb *být (to be)* (být, by, aby, kdyby) are rejected. Lemmata of modal verbs and verbs with similar behaviour are replaced by tag `modal`. These verbs have been found in the list of more than 15 000 verb valencies (Pala and Ševeček P., 1999). In our example it is the verb *zúčastnit* that is removed.

```
být/k5eApFnStMmPaI
by/k5eAp1nStPmCaI
k3xXnSc3
gap
k5eApFnStMmPaP
```

3.2.2 Generalisation of grammatical categories

Exploiting linguistic knowledge, several grammatical categories are not important for verb group description. Very often it is negation (e), or aspect (a - aI stands for imperfectum, aP for perfectum). These categories may be removed. For some of verbs even person (p) can be removed. In our example the values of those grammatical categories have been replaced by `?` and we obtained

```
být/k5e?pFnStMmPa?
by/k5e?p?nStPmCa?
k3xXnSc?
gap
k5e?pFnStMmPa?
```

3.2.3 Finding grammatical agreement constraints

Another situation appears when two or more values of some category are related. In the simplest case they have to be the same – e.g. the value of attribute person (p) in the first and the last word of our example. More complicated is the relation among the values of attribute number (n). They should be the same except when the polite way of addressing occurs, e.g. in *byl byste se jí zúčastnil (you would have participated in it)*. Thus we have to check whether the values are the same or the conditions of polite way of addressing are satisfied. For this purpose we add the predicate `check_num()` that ensures agreement in the grammatical category number and we obtain

```
být/k5e?p_n_tMmPa?
by/k5e?p?n_tPmCa?
k3xXnSc?
gap
k5e?p_n_tMmPa?
check_num(n)
```

3.3 DCG Rules Synthesis

Finally the verb rule is constructed by rewriting the result of the generalisation phase. For the sentence *byla bych se jí zúčastnila (I would have participated in it)* we obtain

```
verb_group(vg(Be,Cond,Se,Verb),
Gaps) -->
be(Be,_,P,N,tM,mP,_),
% být/k5e?p_n_tMmPa?
cond(Cond,_,_,Ncond,tP,mC,_),
% by/k5e?p?n_tPmCa?
{check_num(N,Ncond,Cond,Vy)},
reflex_pron(Se,xX,_,_),
% k3xXnSc?
gap([],Gaps),
% gap
k5(Verb,_,_,P,N,tM,mP,_).
% k5e?p_n_tMmPa?
```

If this rule does not exist in the set of verb rules yet, it is added into it. The meanings of non-terminals used in the rule are following:

`be()` represents auxiliary verb *být*, `cond()` represents various forms of conditionals *by, aby, kdyby*, `reflex_pron()` stands for reflexive pronoun *se (si)*, `gap()` is a special predicate for manipulation with gaps, and `k5()` stands for arbitrary non-auxiliary verb. The particular values of some arguments of non-terminals represent required properties. Simple cases of grammatical agreement are treated through binding of variables. More complicated situations are solved employing constraints like the predicate `check_num()`.

The method has been implemented in Perl. 126 definite clause grammar rules were constructed from the annotated corpus that describe all verb groups that are frequent in Czech.

4 Recognition of Verb Groups

The verb rules have been used for recognition, and consequently for tagging, of verb groups in unannotated text. A portion of sentences which have not been used for learning, has been extracted from a corpus. Each sentence has been ambiguously tagged with LEMMA morphological analyser (Pala and Ševeček, 1995), i.e. each word of the sentence has been assigned to all possible tags. Then all the verb rules were applied to each sentence. The learned verb rules displayed quite good accuracy. For corpus DESAM, a verb rule has been correctly assigned to 92.3% verb groups. We tested, too, how much is this method dependent on the corpus that was used for learning. As the source of testing data we used Prague Tree Bank (PTB) Corpus that is under construction at Charles University in Prague. The accuracy displayed was not different from results for DESAM. It maybe explained by the fact that both corpora have been built from newspaper articles.

Although the accuracy is acceptable for the test data that include also clauses with just one verb, errors have been observed for complex sentences. In about 13% of them, some of compound verb groups were not correctly recognized. It was observed that almost 70% of these errors were caused by incorrect lemma recognition. In the next section we describe a method for fixing this kind of errors.

5 Fixing Misclassification Errors

We combined two approaches, elimination of lemmata which are very rare for a given word form, and inductive logic programming (Popelínský et al., 1999; Pavelek and Popelínský, 1999). The method is used in the post-processing phase to prune the set of rules that have fired for the sentence.

5.1 Elimination of infrequent lemmata

In Czech corpora it was observed that 10% of word positions – i.e. each 10th word of a text – have at least 2 lemmata and about 1% word forms of Czech vocabulary has at least 2 lemmata. (Popelínský et al., 1999; Pavelek and Popelínský, 1999) E.g. word form *při* can be

	correct rules(%)	
	> 1 verb	all
number of examples	349	600
original method	86.8	92.3
+ infrequent lemmata	91.1	94.8
+ ILP	92.8	95.8

Table 2: DESAM: Results for unannotated text

	correct rules(%)	
	> 1 verb	all
number of examples	284	467
original method	87.0	92.1
+ infrequent lemmata	91.6	94.9
+ ILP	92.6	95.5

Table 3: PTB: Results for unannotated text

either preposition *at* (like *at the lesson*) or imperative of *argue*. We decided to remove all the verb rules that recognised a word-lemma couple of a very small frequency in the corpus. Actually those lemmata that did not appear more than twice in DESAM corpus were supposed to be incorrect.

For testing, we randomly chose the set of 600 examples including compound or complex sentences from corpus DESAM. 251 sentences contained only one verb. The results obtained are in Tab. 2. The first line contains the number of examples used. In the following line there are results of the original method as mentioned in Section 4. Next line (+ infrequent lemmata) displays results when the word-lemma couple of a very small frequency have been removed. The column '> 1 verb' concerns the sentences where at least two verbs appeared. The column 'all' displays accuracy for all sentences. Results for corpus PTB are displayed in Tab. 3. It can be observed that after pruning rules that contain a rare lemma the accuracy significantly increased.

5.2 Lemma disambiguation by ILP

Some of incorrectly recognised lemmata cannot be fixed by the method described. E.g. word form *se* has two lemmata, *se* – reflexive pronoun and *s* – preposition *with* and both word-lemma couples are very frequent in Czech. For such cases we exploited inductive logic programming (ILP). The program reads the context of the lemma-ambiguous word and results in disambiguation rules (Popelínský et al., 1999; Pavelek and Popelínský, 1999). We employed ILP system Aleph[1].

Domain knowledge predicates (Popelínský et al., 1999; Pavelek and Popelínský, 1999) have the form `p(Context, first(N),Condition)` or `p(Context,somewhere,Condition)`. where `Context` contains tags of either left or right context (for the left context in the reverse order), `first(N)` defines a subset of `Context`, `somewhere` does not narrow the context. The term `Condition` can have three forms, `somewhere(List)` (the values in `List` appeared somewhere in the define subset of `Context`), `always(List)` (the values appeared in all tags in the given subset of `Context`) and `n_times(N,Tag)` (`Tag` may appear `n`-times in the specified context). E.g. `p(Left, first(2), always([k5,eA]))` succeeds if in the first two tags of the left context there always appears `k5,eA`.

In the last line of Tab. 2 and 3 there is the percentage of correct rules when also the lemma disambiguation has been employed. The increase of accuracy was much smaller than after pruning rules that contain a rare lemma. It has to be mentioned that in the case of PTB about a half of errors (incorrectly recognised verb rules) were caused by incorrect recognition of sentence boundaries.

[1] http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/aleph.html

6 Tagging Verb Groups

We now describe a method for compound verb group tagging in morphologically annotated corpora. We decided to use SGML-like notation for tagging because it allows to incorporate the new tags very easily into DESAM corpus. The beginning and the end of the whole verb group and beginnings and ends of its particular components are marked. For the sentence *byla bych se jí zúčastnila (I would have participated in it)* we receive

```
<vg tag="eApFnStPmCaPr1v0"
fmverb="zúčastnit">
  <vgp>byla</vgp>
  <vgp>bych</vgp>
  <vgp>se</vgp>
  jí
  <vgp>zúčastnila</vgp> </vg>
```

where `<vg> </vg>` point out the beginning and the end of the verb group, `<vgp> </vgp>` mark components (parts) of the verb group. The assigned tag – i.e. values of significant morphologic categories – of the whole group is included as a value of attribute called `tag` in the starting mark of the group. Value of the attribute `fmverb` is the full-meaning verb; this information can be exploited e.g. for searching and processing of verb valencies afterwards. The value of attribute `tag` is computed automatically from the verb rule that describes the compound verb group.

We are also able to detect other properties of compound verb groups. In the example above the new category `r` is introduced. It indicates that the group is reflexive (`r1`) or not (`r0`). The category `v` enables to mark whether the group is in the form of polite way of addressing (`v1`) or not (`v0`). The differences of the `tag` values can be observed comparing the previous and the following examples (*nebyl byste se jí zúčastnil (you would not have participated in it)*)

```
<vg tag="eNpMnStPmCaPr1v1"
fmverb="zúčastnit">
  <vgp>nebyl</vgp>
  <vgp>byste</vgp>
  <vgp>se</vgp>
  jí
  <vgp>zúčastnil</vgp> </vg>
```

The set of attributes can be also enriched e.g. with the number of components. We also plan to include into the attributes of `<vg>` compound verb group type. It will enable to find the groups of the same type but with different word order or the number of components.

7 Discussion

Sometimes compound verb groups are defined in a less general way. Another approach that deal with the recognition and morphological tagging of compound verb groups in Czech appeared in (Osolsobě, 1999). Basic compound verb groups in Czech like active present, passive past tense, present conditional etc., are defined in terms of grammatical categories used in DESAM corpus. Two drawbacks of this approach can be observed. First, verb groups may only be compound of a reflexive pronoun, verbs *to be* and not more than one full-meaning verb. Second, the gap between two words of the particular group cannot be longer than three words. The verb rules defined here are less general then the basic verb groups (Osolsobě, 1999). Actually verb rules make partition of them. Thus we can tag all these basic verb groups without the limitations mentioned above. In contrast to some other approaches we include into the groups also some verbs which are in fact infinitive participants of verb valencies. However, we are able to detect such cases and recognize the "pure" verb groups afterwards. We believe that for some kind of shallow semantic analysis – e.g. in dialogue systems – our approach is more convenient.

We are also able to recognize the form of polite way of addressing a person (which has not equivalent in English, but similar phenomenon appears e.g. in French or German). We extend the tag of a verb group with this information. because it is quite important for understanding the sentence. E.g. in *šel jste (vous êtes allé)* the word *jste* (êtes) should be counted as singular although it is always tagged as plural.

8 Ongoing Research

Our work is a part of the project which aims at building a partial parser for Czech. Main idea of partial parsing (Abney, 1991) is to recognize those parts of sentences which can be recovered reliably and with a small amount of syntactic information. In this paper we deal with recognition and tagging potentially discontinuous verb chunks. The problem of recognition of noun chunks in Czech was addressed in (Smrž and Žáčková, 1998). We aim at an advanced method that should employ a minimum of ad hoc techniques and should be, ideally, fully automatic. The first step in this direction, the method for pruning verb rules, has been described in this paper. In the future we want to make the method even more adaptive. Some preliminary results on finding sentence boundary are displayed below.

In Czech it is either comma and/or a conjunction that make the boundary between two sentences. From the corpus we have randomly chosen 406 pieces of text that contain a comma. In 294 cases the comma split two sentences. All "easy" cases (when comma is followed by a conjunction, it must be a boundary) were removed. It was 155 out of 294 cases. 80% of the rest of examples was used for learning. We used again Aleph for learning. For the test set the learned rules correctly recognised comma as a delimiter of two sentences in 86.3% of cases. When the "easy" cases were added the accuracy increased to 90.8%.

Then we tried to employ this method for automatic finding of boundaries in our system for verb group recognition. Decrease of accuracy was expected but it was quite small. In spite of some extra boundary was found (the particular comma did not split two sentences), the correct verb groups have been found in most of such cases. The reason is that such incorrectly detected boundary splits a compound verb group very rarely.

The last experiment concerned the case when a conjunction splits two sentences and the conjunction is not preceded with comma. There are four such conjunctions in Czech – *a (and), nebo (or), i (even)* and *ani (neither, nor)*. Using Aleph we obtained the accuracy on test data 88.3% for *a* (500 examples, 90% used /for learning) and 87.2% for *nebo* (110 examples). The last two conjunctions split sentences very rarely. Actually, in the current version of corpus DESAM it has never happened.

9 Relevant works

Another approach for recognition of compound verb groups in Czech (Osolsobě, 1999) have been already discussed in Section 7. Ramshaw and Marcus (Ramshaw and Marcus, 1995) views chunking as a tagging problem. They used transformation-based learning and achieved recall and precision rates 93% for base noun phrase (non-recursive noun phrase) and 88% for chunks that partition the sentence. Verb chunking for English was solved by Veenstra (Veenstra, 1999). He used memory-based learner Timbl for noun phrase, verb phrase and propositional phrase chunking.Each word in a sentence was first assigned one of tags I_T (inside the phrase), O_T (outside the phrase) and B_T (left-most word in the phrase that is preceded by another phrase of the same kind), where T stands for the kind of a phrase.

Chunking in Czech language is more difficult than in English for two reasons. First, a gap inside a verb group may be more complex and it may be even a whole sentence. Second, Czech language is a free word-order language what implies that the process of recognition of the verb group structure is much more difficult.

10 Conclusion

We described the method for automatic recognition of compound verb groups in Czech sentences. Recognition of compound verb groups was tested on unannotated text randomly chosen from two different corpora and the accuracy reached 95% of correctly recognised verb groups. We also introduced the method for automatic tagging of compound verb groups.

Acknowledgements

We thank to anonymous referees for their comments. This research has been partially supported by the Czech Ministry of Education under the grant VS97028.

References

S. Abney. 1991. Parsing by chunks. In *Principle-Based Parsing*. Kluwer Academic Publishers.

N. M. Marques, G. P. Lopes, and C. A. Coelho. 1998. Using loglinear clustering for subcategorization identification. In *Principles of Data Mining and Knowledge Discovery: Proceedings of PKDD'98 Symposium, LNAI 1510*, pages 379–387. Springer.

K. Osolsobě. 1999. Morphological tagging of composed verb forms in Czech corpus. Technical report, Studia Minora Facultatis Philosophicae Universitatis Brunensis Brno.

K. Pala and Ševeček P. 1999. Valencies of Czech Verbs. Technical Report A45, Studia Minora Facultatis Philosophicae Universitatis Brunensis Brno.

K. Pala and P. Ševeček. 1995. Lemma morphological analyser. User manual. Lingea Brno.

K. Pala, P. Rychlý, and P. Smrž. 1997. DESAM - annotated corpus for Czech. In *In Plášil F., Jeffery K.G.(eds.): Proceedings of SOFSEM'97, Milovy, Czech Republic. LNCS 1338*, pages 60–69. Springer.

T. Pavelek and L. Popelínský. 1999. Mining lemma disambiguation rules from Czech corpora. In *Principles of Knowledge Discovery in Databases: Proceedings of PKDD'99 Conference, LNAI 1704*, pages 498–503. Springer.

L. Popelínský, T. Pavelek, and T. Ptáčník. 1999. Towards disambiguation in Czech corpora. In *Learning Language in Logic: Working Notes of LLL Workshop*, pages 106–116. JSI Ljubljana.

L. A. Ramshaw and M. P. Marcus. 1995. Text chunking using transformation-based learning. In *Proceedings of the Third ACL Workshop on Very Large Corpora*. Association for Computational Linguistics.

P. Smrž and E. Žáčková. 1998. New tools for disambiguation of Czech texts. In *Text, Speech and Dialogue: Proceedings of TSD'98 Workshop*, pages 129–134. Masaryk University, Brno.

J. Veenstra. 1999. Memory-based text chunking. In *Machine Learning in Human Language Technology: Workshop at ACAI 99.*

E. Žáčková and K. Pala. 1999. Corpus-based rules for Czech verb discontinuous constituents. In *Text, Speech and Dialogue: Proceedings of TSD'99 Workshop, LNAI 1692*, pages 325–328. Springer.

E. Žáčková and L. Popelínský. 2000. Automatic tagging of compound verb groups in Czech corpora. In *Text, Speech and Dialogue: Proceedings of TSD'2000 Workshop, LNAI.* Springer.