Automatic Tagging of Compound Verb Groups in Czech Corpora*

Eva Žáčková, Luboš Popelínský and Miloslav Nepil

NLP Laboratory, Faculty of Informatics, Masaryk University
Botanická 68, CZ-602 00 Brno, Czech Republic
E-mail: {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 loose their original meaning. We present an improved method for automatic synthesis of verb rules that describe all compound verb groups frequent in Czech. Using these rules we can find compound verb groups in unannotated texts with high accuracy. The system for tagging compound verb groups in an annotated corpus that exploits the verb rules is described.

Keywords: compound verb groups, corpora, 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 (semantic) 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 [2], 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. In the sentence above the word *jsem* is tagged as a verb in present tense, but the whole verb group to which it belongs – *jsem nevěděla (I did not know)* – is in the

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

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.

We consider a compound verb group to be a list of cohesive verbs and maybe reflexive pronouns *se, si*. Such a group is obviously compound of auxiliary and full-meaning verbs, e.g. *budu se smát* where *budu* is auxiliary verb (like *will* in English), *se* is the reflexive pronoun and *smát* means *to laugh*. 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 characterize the verb group as a whole. 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. In [4] a proposal of the method for automatic finding of compound verb groups in the corpus DESAM is introduced. We describe here an 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 their tags. We describe also some improvements that allow to increase the accuracy of verb group recognition.

2 Learning Verb Rules

The process of learning verb rules [4] is split into three steps: finding verb chunks (i.e. finding boundaries of simplex sentences in compound or in complex sentences, and elimination of gap words), generalization of the chunk and verb rule synthesis.

2.1 Verb Chunks

The observed properties of a verb group are the following: their components are either verbs or reflexive pronoun *se* (*si*); a verb group cannot cross 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 simplex 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 mentioned in the first section (*byla bych se jí zúčastnila (I would have participated in it)*):

```
být/k5eApFnStMmPaI by/k5eAp1nStPmCaI si/k3xXnSc4
on/k3xPgFnSc2p3 zúčastnit/k5eApFnStMmPaP
```

After substitution of gaps we obtain

```
být/k5eApFnStMmPaI by/k5eAp1nStPmCaI si/k3xXnSc4
gap
zúčastnit/k5eApFnStMmPaP
```

2.2 Generalization

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

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. In our example it is the verb *zúčastnit* that is removed.

```
být/k5eApFnStMmPaI by/k5eAp1nStPmCaI k3xXnSc4
gap k5eApFnStMmPaP
```

Generalization of grammatical categories. By exploiting linguistic knowledge we can observe that some grammatical categories are not important for verb group description. These categories may also be removed. Very often it is negation (e), or aspect (a). For some of verbs even person (p) can be removed. Values of those grammatical categories have been replaced by `?` in our example.

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

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)
```

2.3 DCG Rules Synthesis.

Finally the verb rule is constructed by rewriting the result of the generalization phase.

```
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?
```

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 the reflexive pronoun *se (si)*, `gap()` is a special predicate for manipulation with gaps, and `k5()` stands for an 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 by 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. Although the method described displays quite good accuracy on the test set (the correct verb rule has been corrrectly assigned to 92.3% verb groups), it has some drawbacks. In about 13% of complex sentences, some of compound verb groups were not correctly recognized. It was observed that almost 46% of errors were caused by incorrect lemma recognition.

3 Fixing Misclassification Errors

To fix the most frequent error – incorrectly recognised lemma of a word in the verb group – we combined two approaches. Rules that recognised a word-lemma couple of a very small frequency in the corpus (actually those that did not appear there) were supposed to be incorrect. The second approach that exploits inductive logic programming (ILP) techniques reads the context of the lemma-ambiguous word and results in disambiguation rules [3].

For testing we randomly chose the set of 600 examples including compound or complex sentences. 251 sentences contained only one verb. The results obtained are in Tab. 1. The first line contains the number of examples used. In the following line there are results for the original method. Then the results with the word-lemma couple of a very small frequency have been removed; the accuracy is in the third line. In the last line is the percentage of correct rules after the lemma disambiguation was employed. The column '> 1 verb' concerns the sentences where at least two verbs appeared. The last column displays accuracy for all sentences.

4 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

Table 1. Fixing

	number of correct rules(%) > 1 verb	all
number of examples	349	600
original method	86.8	92.3
− not frequent lemmata	91.1	94.8
+ lemma dis. employed	92.8	95.8

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.

5 Conclusion

Let us notice that sometimes compound verb groups are defined in a less general way. In contrast to these 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 dialog 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). As it is quite important for understanding the sentence we extend the tag of a verb group with this information. E.g. in *šel jste (vous êtes allé)* the word *jste* should be counted as singular although it is always tagged as plural.

Another approach that deals with the recognition and morphological tagging of compound verb groups in Czech appeared in [1]. 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 [1]. Actually verb rules make partition of them. Thus we can tag all the basic verb groups without the limitations mentioned above.

References

1. K. Osolsobě. Morphological Tagging of Compound Verb Forms in Czech Corpus. Technical report, Studia Minora Facultatis Philosophicae Universitatis Brunensis Brno, 1999.
2. K. Pala, P. Rychlý, and P. Smrž. DESAM – Annotated Corpus for Czech. In *Plášil F., Jeffery K.G.(eds.): Proceedings of SOFSEM'97, Milovy, Czech Republic. LNCS 1338*, pages 60–69. Springer, 1997.
3. T. Pavelek and L. Popelínský. 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, 1999.
4. E. Žáčková and K. Pala. Corpus-Based Rules for Czech Verb Discontinuous Constituents. In *Text, Speech and Dialogue: Proceedings of TSD'99 Workshop, LNAI 1692*, pages 325–328. Springer, 1999.