Extraction of Syntactic Structures Based on the Czech Parser Synt

Miloš Jakubíček

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
xjakub@fi.muni.cz

Abstract. In this paper we describe the usage of the syntactic parser `synt` (developed in the NLP Centre at Masaryk University) to gain information about syntactic structures (such as noun or verb phrases) of common sentences in Czech. These structures are from the analysis point of view usually identical to nonterminals in the grammar used by the parser to find possible valid derivations of the given sentence. The parser has been extended in such a way that enables its highly ambiguous output to be used for extracting those syntactic structures *unambiguously* and gives several ways how to identify them. To achieve this, some previously unused results of syntactic analysis have been evolved leading to more precise morphological analysis and hence also deeper distinction among various syntactic (sub)structures. Finally, we present an application for shallow valency extraction.

1 Introduction

Usually, a derivation tree is presented as the main output of syntactic parsing of natural languages, but currently most of the syntactic analysers for Czech lack precision, i.e. there are more (actually, in some cases up to billions) trees given on the output. However there are many situations in which it is not necessary and sometimes even not desirable to have such derivation trees, may it be information extraction and retrieval, transformation of sentences into a form of predicate(arguments) or shallow valency extraction. In such cases we rather need to process whole syntactic structures in the given sentence, especially noun, prepositional and verb phrases, numerals or clauses. Moreover, so as not to end up with the same problems as with the standard parser output, we need to identify the structures unambiguously.

Therefore we modified the Czech parser `synt` so that it is possible to gain syntactic structures corresponding to the given nonterminal in a number of ways according to the user's choice. To improve the structures detection, we also employed the results of contextual actions used in `synt` as described in Section 4, which increased the precision of morphological analysis by almost 30 %. We also present results of the extraction from sample sentences as well as the usage for shallow valency extraction from annotated corpora.

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008, pp. 56–62, 2008. © Masaryk University, Brno 2008

2 Syntactic parser synt

Syntactic parser synt [1] has been developed for several years in the Natural Language Processing Centre at Masaryk University. It performs a chart-type syntactic analysis based on the provided context-free head-driven phrase-structure grammar for Czech. For easy maintenance, this grammar is edited in form of a metagrammar (having about 200 rules) from which the full grammar can be automatically derived (having almost 4,000 rules). Contextual phenomena (such as case-number-gender agreement) are covered using the per-rule defined contextual actions.

In recent measures [2, p. 77] it has been shown that synt accomplishes a very good recall (above 90 %) but the analysis is highly ambiguous: for some sentences even billions of output syntactic trees can occur. There are two main strategies developed to fight such ambiguity: first, the grammar rules are divided into different priority levels which are used to prune the resulting set of output trees. Second, every grammar rule has a ranking value assigned from which the ranking for the whole tree can be efficiently computed in order to sort the trees on the output accordingly.

For the purpose of the extraction, the internal parsing structure of synt is used, the so called *chart*, an acyclic multigraph which is built up during the analysis holding all the resulting trees. What is important about chart is its polynomial size [3, p. 133] implying that it is a structure suitable for further effective processing – as the number of output trees can be up to exponential regarding to the length of the input sentence, processing of each tree separately would be otherwise computationally infeasible. By processing of the chart we refer to the result of the syntactic analysis, i.e. to the state of the chart after the analysis.

3 Extraction of structures

Several ways how to identify the given syntactic structures have been developed respecting the (from the nature of language given) reality that these structures differ a lot in their inner form and thus no universal procedure can be used for all of them. Since we want the output of the extraction to be unambiguous, the extraction covers all possible structures and their combination that result from the analysis. There are two very straightforward approaches for structures detection which consist in extracting the biggest or smallest found structure, however to achieve quality results, more sophisticated methods have to be employed for each structure/nonterminal separately. Speaking about biggest or smallest we mean that regarding to the fact that many of the rules in the grammar used by synt are recursive. The results for various nonterminals are listed in Examples 1–4.

- *Example 1.* – clause (nested)
  Input:

Muž, který stojí u cesty, vede kolo.
(A man who stands at the road leads a bike.)
Output:
`[0-9): Muž , , vede kolo` *(a man leads a bike)*
`[2-6): který stojí u cesty` *(who stands at the road)*

- *Example 2.* – verb phrase
Input:
Kdybych to byl býval věděl, byl bych sem nechodil.
(If I had known it, I would not have come here.)
Output:
`[0-5)`[1] `: byl býval věděl` *(had known)*
`[6-10): byl bych nechodil` *(would not have come)*

- *Example 3.* – clause (sequence)
Input:
Vidím ženu, která drží růži, která je červená.
(I see a woman who holds a flower which is red.)
Output:
`[0-3): Vidím ženu ,` *(I see a woman)*
`[3-7): která drží růži ,` *(who holds a flower)*
`[7-10): která je červená` *(which is red)*
- *Example 4.* – noun phrase
Input:
Tyto normy se však odlišují nejen v rámci různých národů a států, ale i v rámci sociálních skupin, a tak považuji dřívější pojetí za dosti široké a nedostačující.
(But these standards differ not only within the scope of various nations and countries but also within the scope of social groups and hence I consider the former conception to be wide and insufficient.)
Output:
`[0-2)`: Tyto normy *(These standards)*
`[6-12)`: v rámci různých národů a států *(within the scope of various nations and countries)*
`[15-19)`: v rámci sociálních skupin *(within various social groups)*
`[23-30)`: dřívější pojetí za dosti široké a nedostačující *(former conception for wide and insufficient)*

4 Morphological refinement

In order to further divide big structures into separate meaningful segments it is possible to part them according to the morphological agreement – i.e. in such a way that words in each structure agree in case, number and gender. To

[1] The numbering denotes a (left inclusive, right exclusive) range of the structure in the input sentence (i.e. words indices).

improve this technique some previously unused results of the syntactic analysis have been involved, namely the contextual actions used by the parser to handle the case-number-gender agreement. In each analysis step, the results of the contextual actions are propagated bottom-up so that they can be used in the next step to prune possible derivations.

Table 1. A comparison of morphological tagging before and after the refinement. The whole sentence in English was: *There was a modern shiny car standing on a beautiful long street.* Note that for readability purpose we abbreviate the tags so that `k7{c4,c6}` stands for `k7c4, k7c6`.

word	**before**	**after**
Na *(on)*	k7{c4, c6}	k7c6
krásné *(beautiful)*	k2eA{gFnPc1d1, gFnPc4d1, gFnPc5d1, gFnSc2d1, gFnSc3d1, gFnSc6d1, gInPc1d1, gInPc4d1, gInPc5d1, gInSc1d1wH, gInSc4d1wH, gInSc5d1wH, gMnPc4d1, gMnSc1d1wH, gMnSc5d1wH, gNnSc1d1, gNnSc4d1, gNnSc5d1}	k2eAgFnSc6d1
dlouhé *(long)*	k2eA{gFnPc1d1, gFnPc4d1, gFnPc5d1, gFnSc2d1, gFnSc3d1, gFnSc6d1, gInPc1d1, gInPc4d1, gInPc5d1, gInSc1d1wH, gInSc4d1wH, gInSc5d1wH, gMnPc4d1, gMnSc1d1wH, gMnSc5d1wH, gNnSc1d1, gNnSc4d1, gNnSc5d1}	k2eAgFnSc6d1
ulici *(street)*	k1gFnSc3, k1gFnSc4, k1gFnSc6	k1gFnSc6
stálo *(stand)*	k5eAaImAgNnSaIrD	k5eApNnStMmPaI[2]
moderní *(modern)*	k2eA{gFnPc1d1, gFnPc4d1, gFnPc5d1, gFnSc1d1, gFnSc2d1, gFnSc3d1, gFnSc4d1, gFnSc5d1, gFnSc6d1, gFnSc7d1, gInPc1d1, gInPc4d1, gInPc5d1, gInSc1d1, gInSc4d1, gInSc5d1, gMnPc1d1, gMnPc4d1, gMnPc5d1, gMnSc1d1, gMnSc5d1, gNnPc1d1, gNnPc4d1, gNnPc5d1, gNnSc1d1, gNnSc4d1, gNnSc5d1}	k2eAgNnSc1d1, k2eAgNnSc4d1, k2eAgNnSc5d1
nablýskané *(shiny)*	k2eA{gFnPc1d1rD, gFnPc4d1rD, gFnPc5d1rD, gFnSc2d1rD, gFnSc3d1rD, gFnSc6d1rD, gInPc1d1rD, gInPc4d1rD, gInPc5d1rD, gInSc1d1wHrD, gInSc4d1wHrD, gInSc5d1wHrD, gMnPc4d1rD, gMnSc1d1wHrD, gMnSc5d1wHrD, gNnSc1d1rD, gNnSc4d1rD, gNnSc5d1rD}	k2eAgNnSc1d1, k2eAgNnSc4d1, k2eAgNnSc5d1
auto *(car)*	k1gNnSc1, k1gNnSc4, k1gNnSc5	k1gNnSc1, k1gNnSc4, k1gNnSc5

So far these outcomes in form of morphological values have not been used in any other way. Our enhancement backpropagates these values after the analysis top-down to the chart nodes, i.e. input words, and prunes their original morphological tagging. This leads to more precise morphological analysis and hence it also enables more exact distinction between substructures according to grammar agreement. A detailed example of the impact of morphological refinement on particular sentence is provided in Table 1.

Testing on nearly 30,000 sentences from Czech annotated corpus DESAM [4] has shown that it is possible to increase the number of unambiguously analysed

[2] The inconsistence in tagging on this row has purely technical background – the tag set has been changed.

words by almost 30 % using this method while the number of errors introduced consequently remains very low, as shown in Table 2.

Table 2. Morphological refinement results on the DESAM corpus.

value	**before**	**after**
average unambiguous words	20,733 %	46,1172 %
average pruned word tags	38,3716 %	
error rate [3]	< 1,46 %	
number of sentences	29 604	

Parting structures according to their grammatical agreement is useful, for example, when extracting noun or prepositional phrases, as can be seen in Example 5 (compare with Example 4 where the same sentence is extracted without morphological parting).

Example 5.
Input:
Tyto normy se však odlišují nejen v rámci různých národů a států, ale i v rámci sociálních skupin, a tak považuji dřívější pojetí za dosti široké a nedostačující.
(But these standards differ not only within the scope of various nations and countries but also within the scope of social groups and hence I consider the former conception to be wide and insufficient.)
Output:

```
[0-4): Tyto normy se však
```
(But these standards)

```
[6-8): v rámci
```
(within the scope)

```
[8-12): různých národů a států
```
(various nations and countries)

```
[13-17): ale i v rámci
```
(but also within the scope)

```
[17-19): sociálních skupin
```
(social groups)

```
[23-25): dřívější pojetí
```
(former conception)

```
[25-30): za dosti široké a nedostačující
```
(for wide and insufficient)

Specific modifications how to extract nonterminals with important semantical representation have been developed. Furthermore, these settings can be extended to other (possibly new) nonterminals easily as they are available as command-line parameters.

[3] As an error we consider a situation when the correct tag has been removed during the refinement process. Actually the error rate is even lower since many of the results marked as wrong were caused by an incorrect tag in the corpus.

5 Applications: shallow valency extraction

Currently, a new verb valencies lexicon for Czech, called *Verbalex* [5], is being developed in the NLP Centre. As building of such a lexicon is a very time-consuming long-term task for linguists professionals, it is extremely important to use any possibilities to make this process easier for them. Therefore, we extended the extraction of structures so that it performs a shallow valency extraction from annotated corpora. The main idea is as follows: first, we extract separate clauses, then we identify individual verbs or verb phrases and finally we find noun and prepositional phrases within each clause. Sample output in BRIEF format is provided in Example 6. Moreover, such basic extraction might be used for approximating the coverage of the valency lexicon by finding verbs that are not included there.

Example 6.

```
 ; extracted from sentence: Nenadálou finanční krizi musela
podnikatelka řešit jiným způsobem .
řešit <v>hPTc4,hPTc7
```

(The businessman had to ***solve*** *the sudden financial crisis in another way.)*

```
; extracted from sentence: Hlavní pomoc ale nacházela v dalších
obchodních aktivitách .
nacházet <v>hPTc4,hPTc6r{v}
```

(However she ***found*** *the main help in further business activities.)*

```
; extracted from sentence: U výpočetní techniky se pohybuje
v rozmezí od 8000 Kč do 16000 Kč .
pohybovat <v>hPTc2r{u},hPTc6{v}
```

(By IT [it] ***ranges*** *between 8,000 Kč and 16,000 Kč.)*

6 Conclusions

We presented recent improvements in the Czech parser synt that can be used for extracting various syntactic (sub)structures. We also showed practical usage of syntactic analysis for refining morphological tagging as well as examples using the resulting tagging for structures distinction. Furthermore, we presented an application of structures extraction, namely shallow extraction of valencies.

In the future there will be further work on the development of this extraction. We would like to compare the results of morphological refinement with similar oriented methods (e.g. with morphological disambiguation as described in [6]) as well as perform more detailed experiments with the shallow valency extraction on big annotated corpora.

Acknowledgements. This work has been partly supported by the Academy of Sciences of Czech Republic under the project 1ET100300414 and by the Ministry of Education of CR within the Center of basic research LC536 and in the National Research Programme II project 2C06009.

References

1. Kadlec, V., Horák, A.: New meta-grammar constructs in Czech language parser synt. In: Proceedings of TSD 2005, LNCS 3658, Springer-Verlag (2005), pp. 85–92.
2. Kadlec, V.: Syntactic analysis of natural languages based on context-free grammar backbone. Ph.D. thesis, Faculty of Informatics, Masaryk University, Brno (2007).
3. Horák, A.: The Normal Translation Algorithm in Transparent Intensional Logic for Czech. Ph.D. thesis, Faculty of Informatics, Masaryk University, Brno (2001).
4. Pala, K., Rychlý, P., Smrž, P.: DESAM – Annotated Corpus for Czech. In: Proceedings of SOFSEM '97, Springer-Verlag (1997) 523–530.
5. Hlaváčková, D., Horák, A., Kadlec, V.: Exploitation of the Verbalex verb valency lexicon in the syntactic analysis of Czech. In: Proceedings of TSD 2006, Brno, Springer-Verlag (2006), pp. 85–92.
6. Šmerk, P.: Unsupervised learning of rules for morphological disambiguation. In: Proceedings of TSD 2004, Brno, LNCS 3206, Springer-Verlag (2004), pp. 211–216.