M. Hnátková, V. Petkevič, H. Skoumalová

LINGUISTIC ANNOTATION OF CORPORA IN THE CZECH NATIONAL CORPUS[1]

0. Introduction

In the project *Czech National Corpus and the Corpora of Other Languages* the key role is played by extensive corpora (comprising hundreds of millions of words) of written contemporary Czech: SYN2005, SYN2006PUB, SYN2009PUB, SYN2010, SYN (cf. **http://www.korpus.cz**). One of the important characteristics of these corpora is the fact that their texts are lemmatized and morphologically annotated. We shall describe the whole system and individual phases of an entirely automatic process of linguistic annotation.

1. Phases of linguistic annotation

The whole linguistic annotation consists of the following three phases:

a) morphological phase
b) disambiguation phase
c) complementary phase.

1.1 Morphological phase

At the beginning of the processing the text of a document is formed by a sequence of characters including the blank spaces. The morphological phase is composed of the following parts:
a) "premorphological" phase: preprocessing of the input plain text consisting in the concatenation of neighbouring strings, or in splitting strings into several strings separated by blanks, as well as in corrections of obvious typos;

[1] This paper was supported by the grant MSM0021620823 of the Ministry of Education of the Czech Republic.

b) morphological analysis in a broader sense – it involves:

- *tokenization*: identification of individual textual words and punctuation as independent elements – *tokens*;
- *sentence segmentation*: separation of the input text into sentences based on punctuation and segmentation rules;
- *morphological analysis proper*: each token is assigned: a) all of its lemmas, i.e. representations of lexemes pertaining to the token; b) all of its part-of-speech (POS) and morphological properties in the form of *tags*.

Morphological analysis is realized by a *morphological analyzer*: it analyses every token and assigns to it all of its lemmas and POS properties regardless of context, i.e. only on the basis of the token itself and its properties contained in the morphological dictionary. The dictionary contains ca 350.000 lexemes including 155.000 proper names: it includes primarily the vocabulary of the *Dictionary of Standard Czech Language*[1] and the *Dictionary of Standard Czech*[2] which contain 194.000 and 57.000 lexemes, respectively. In addition, it comprises also some other lexemes derived from these lexemes: e.g. deadjectival nouns and adverbs. Moreover, the morphological lexicon is gradually extended.

c) „postmorphological" phase: it consists in ad hoc corrections of possible errors in morphological analysis that – for organizational reasons – could not be rectified in the morphological dictionary.

1.2 Disambiguation phase

The morphological phase is followed up with the *disambiguation one*: homographs are subject to the disambiguation of lemmas and POS and morphological tags of individual tokens. The natural language texts can generally be POS and morphologically disambiguated by the three possible types of methods:

---

[1] Slovník spisovného jazyka českého. Praha, 1960–1971.

[2] Slovník spisovné češtiny. Praha, 1994.

(i) *statistically* (*stochastically*) – on the basis of machine learning;
(ii) by linguistic *rules*: either (ii1) rules automatically inferred from texts, or (ii2) hand-crafted rules;
(iii) cooperation of type (i) a (ii) – a *hybrid* method.

For the disambiguation of Czech corpora the method (iii) was selected as optimum: it includes the statistical tagger called *MorČe* (=Morphologie Češtiny) and the *LanGr* tagger based on hand-crafted rules (of the (ii2) type).

The statistical tagger *MorČe* is based on machine-learning: it uses a training corpus of several hundreds of thousands of words; in addition, some of its *features* or their combination can be parameterized. At present, it is the best statistical tagger of Czech. It is very robust: it need not be retrained in case the tagset or input data is moderately modified.

The other tagger called *LanGr* is based on a system of thousands of manually written rules that are (a) developed on the basis of linguistic introspection and checked on corpus data, and (b) non-automatically inferred from corpus data. Linguistic rules of the *LanGr* tagger are written in a special programming language and their performance consists in context-based gradual deletion of incorrect lemmas and tags assigned to individual tokens. First, the tagger processes the output of morphological analysis which assigns every token all of its tags and lemmas. After morphological analysis, in a typical case every token in the input sentence is assigned a lemma and tag that are correct in the given context, i.e. the *recall* is generally almost 100%, for the morphological dictionary includes the whole vocabulary of contemporary Czech. However, as the morphological analyzer assigns all tokens all of its lemmas and tags regardless of the context, the tokens are assigned the highest amount of incorrect tags. This fact is quantified by the *precision* measure: it is lowest possible on disambiguation input. The disambiguation consists in keeping the best recall as possible (close to 100%) and in increasing precision by removing lemmas and tags incorrect in the given context. Specifically,

in some cases morphological analysis also yields some very general tags that are transformed to more specific tags during the disambiguation.

Example

(1) *Zajímá mě* ***poslech*** *rozhlasu.*
E. lit. *Interests me* ***listening*** *of radio*.
E. *I am interested in* ***listening*** *to the radio*.

Morphological analysis assigns the word *poslech* the following 4 pairs (lemma, tag):
a) lemma="poslech", NounMascInanNomSg, tag=NNIS1-----A-----
(E. lemma: *listening*)
b) lemma="poslech", NounMascInanAccSg, tag=NNIS4-----A-----
(E. lemma: *listening*)
c) lemma="posel", NounMascAnimatLocPl, tag=NNMP6-----A-----
(E. lemma: *messenger*)
d) lemma="poslechnout", VrbPastMascSgAct,
tag=VpYS---XR-AA---6 (E. lemma: *obey*)

The task of the disambiguation is to remove incorrect lemmas and tags b), c) and d), since the only correct tag is the a) alternative: i.e. nom. sg. masc. inanimate of the noun lemma="poslech". Disambiguation rules are contained in two groups: a) safe rules, b) heuristic rules; they are applied to the input sentence tokens and remove their tags and lemmas that are contextually inappropriate (e.g. tags NNIS4-----A----- a NNMP6-----A----- and corresponding lemmas of the token *poslech* in sentence (1)). An input sentence is more and more disambiguated by the rules' application until – ideally – full disambiguation is achieved, i.e. each token is assigned the only correct lemma and tag, i.e. the a) alternative of *poslech* in sentence (1). In case the rule-based tagger is unable to entirely disambiguate all tokens of an input sentence, i.e. some lemmas are still assigned more tags, the remaining incorrect ones are removed by the statistical tagger *MorČe*.

The POS and morphological disambiguation also involves the collocational module *Phras* identifying and properly disambiguating

so-called grammatical and non-grammatical collocations. Thus, the following modules take part in the disambiguation process:

(i) *LanGr* tagger based on hand-crafted rules;

(ii) collocational/phraseme *Phras* module based on manually written rules and dictionary of collocations;

(iii) parameterizable stochastic tagger *MorČe*.

The collaboration of the modules can be described by the following sequence of operations applied to a sentence:

1th step: the output of morphological analysis is processed by safe rules. The rules gradually disambiguate the sentence, i.e. the number of incorrect tags decreases. The process continues till there is nothing to disambiguate, i.e. till the rules in recurrent cycles exhaust their disambiguation capacity;

2nd step: *Phras*, the collocational module is invoked: it identifies the collocations in the sentence and performs their disambiguation;

3rd step: both safe and heuristic rules of the *LanGr* tagger are applied till there is nothing to disambiguate;

4th step: the remaining incorrect tags untouched by the *LanGr* tagger are removed by the stochastic tagger *MorČe*.

Our experience shows that this is the optimum disambiguation strategy for such a morphologically complex language as Czech. This is due to the following main properties of the language system of Czech which considerably influence the accuracy of the disambiguation of Czech sentences: a) complex morphology (number of tags is ca 5000, out of which ca 1500 are really exploited); b) high morphological syncretism; c) high amount of casual, synchronically unmotivated ambiguity; d) many exceptions and irregularities in morphology and syntax; e) relatively free word-order enabled by the a) property above; e) relatively few reference points in a sentence that could be safely exploited by disambiguation rules; f) strict rules of orthography including the punctuation ones.

1.3 Complementary phase

Following the two main phases comes the third, complementary one. It consist of the following steps:
(a) involvement of the "aspect module", which assigns the verbs an aspect value (in future, the aspect assignment will be performed within morphological analysis). In Czech the verbs can be:

- perfective (e.g. *přidat*, E. add, R. добавить)
- imperfective (e.g. *přidávat*, E. add, R. добавлять)
- biaspectual (e.g. *adaptovat*, E. adapt, R. адаптировать).

(b) inclusion of various parameterizable modules:
(b) correction of tokenization based on already disambiguated tokens;
(c) optional POS corrections (e.g. adverb ↔ particle);
d) optional tagging of some morphosyntactic functions (auxiliary verbs) etc.

2. Evaluation

The accuracy (recall+precision) of the whole system is close to 95%, the precision of individual steps was not counted. The recall of morphological analysis is 99.25%, the recall of safe disambiguation rules is 99.09%, safe rules + *Phras* module have the recall = 99.07%; with heuristic rules added the rule-based system has recall = 98.82%. The big gap between 98.82% and 95% is due to the complexity of the remaining disambiguation problems solved by the statistical tagger *MorČe* which, however, considerably increases precision.