Morphological Guesser of Czech Words*

Jaroslava Hlaváčová

Institute of the Czech National Corpus, Faculty of Arts, Charles university, nám. J. Palacha 2, 116 38 Prague 1, Czech Republic
jaroslava.hlavacova@ff.cuni.cz

Abstract. If a corpus is submitted to a morphological analysis, there always remain some words that the analyser could not recognize (foreign names, misspellings, ...). However, if a human reads the texts, he usually understands them, even if he does not know as many words as there are in the lexicon used by the morphological analyser. The language itself helps him to recognize unknown words. It is not only semantics or syntax but also pure morphology of unknown words that can contribute to their understanding.

In this article, I describe a “guesser” that can lower the amount of unrecognized words after the “classical” morphological analysis of the Czech texts. It was tested on the Czech National Corpus.

1 Introduction

The automatic morphological analysis of Czech texts is processed on the basis of a large lexicon (about the morphological analysis of the Czech see [2]). However, there are still quite a lot of words in texts of the Czech National Corpus (CNC) that are not recognized by the morphological analyser.

A relatively great amount of unrecognized words in the Czech National Corpus (approximately 2.3 %) causes a lot of problems during the morphological and morphosyntactic disambiguation of Czech sentences. That’s why I tried to find a method, which would help to lower this rate. In fact, for the disambiguation (and for many other analyses) we do not need to recognize the words fully. It is sufficient to recognize the values of some of their morphological categories. Sometimes only part of speech will do. In other words we need a guesser that would assign unknown words as many morphological values as possible. (About a guesser for French – see [3].)

The Czech has a very nice feature - morphology of a word is “visible” by its ending. It can be of various lengths, but generally, the last four letters of a word are in a great majority of cases sufficient for guessing the most important morphological values – see [1].

On the other hand, the Czech has a very ugly feature – many word forms are ambiguous. One word form can have not only 2, 3, 4 or for some words even 5 parts of speech, but even many more combinations of case, number, gender, etc. That’s why the morphological analysis assigns to many words more than one tag.

* This research was supported by the GACR, Grant Nr. 405/96/K214.
2 The Guesser

To construct the guesser I try to employ as many specific features of Czech words as possible. The most important one has been already mentioned - the last four letters of a word form are usually enough to recognize its basic morphological categories, namely part of speech, gender, number, case, verbal tense and person, and for specific parts of speech some other. This feature is common to more Slavonic languages.

We need to have the complete list of all possible endings together with the tags that can be assigned to them. Generally, it is not easy to determine, where the (morphological) ending of an unknown word begins. That’s why I will call these endings segments in the following sections. The term segment is borrowed from another project – MOZAIKA (see [1]), which dealt with a similar problem. I have to stress that segments need not coincide with the morphological endings of words. If I needed to refer to an end of a word as a part of the word, I will call it tail to distinguish it from its morphological ending.

To summarize the terminology I am using:

**Ending** is a morphological ending of a word form (e.g. ing in the word using, many English words have none, e.g. problem).

**Segment** is a string of letters that can occur at the end of a word form (e.g. g, ng, ing, sing - they can occur at the end of many words) together with all the possible morphological tags assigned to them.

**Tail** is an end string of a particular word (e.g. e, ge, age, uage for the word language).

I take into consideration segments and tails of the length 2, 3 or 4 only. Segments of the length 1 - a single letter - usually have too many possible tags.

For building the list of segments we would need a list of all Czech word forms, perfectly morphologically tagged. Such a list does not exist. However, the Czech National Corpus is big enough (100 mil. words) for us to obtain a very large list of segments, so I used it.

I have taken all the recognized words of the corpus. For every word longer than 4 letters I took its 4-letter tail together with the set of all the morphological tags that the analyser assigned to the word. I obtained the file of segments. I grouped all these segments together. The resulting file contains for every segment all the possible tags that were found in the corpus.

If the corpus is big and representative enough, all the combinations are covered. As we will never be sure that the list is complete (and it will never be so), we should continually add all the new combinations that will occur during the gathering of texts. The contribution of every new text is smaller and smaller, if we reach a reasonably extensive file of segments - see Figure 1. During creation of the list of segments from the Czech National Corpus I registered size of the list after addition of every new text. The result is a curve depicting speed of growth of the number of segments. At the horizontal axis there is number of words processed, at the vertical one total number of segments created on the basis of the words processed. I used a set of approximately 10 000 segments as the “nest-egg”. It is visible, that number of segments increases steeply from the beginning (it would be even better visible, if I used a smaller nest-egg). As the number of words