Morphological Tagging of the Brno Spoken Corpus

Dana Hlaváčková[1] and Radek Sedláček[2]

[1] Faculty of Arts, Masaryk University Brno
Arna Nováka 1, 660 88 Brno, Czech Republic
dana.hl@email.cz
[2] Faculty of Informatics, Masaryk University Brno
Botanická 68a, 602 00 Brno, Czech Republic
rsedlac@fi.muni.cz

Abstract. In this paper, we deal with the morphological tagging of the spoken corpus. First, we present the content of the corpus – recording dialogues and their transcription. Next, the process of the morphological tagging of such a corpus is described. We pay attention to the necessary adjustments of the program used for the tagging. These adjustments are based on the description of the regular differences between the written form of the Czech language and the spoken one. Finally, we discuss specific features of the spoken language and estimate the percentage of the successfully tagged word forms.

1 Introduction

The Brno Spoken Corpus has been built at the Faculty of Arts in cooperation with the Faculty of Informatics of Masaryk University in Brno since 1996. The corpus consists of the tape recordings of speech that were transcribed into a text editor. At present, the corpus includes 250 text files, i.e. about 400,000 word occurrences. Concordances (showing where and how often the word is used in the corpus) may be found with the support of GCQP (Graphical Corpus Query Processor) [1]. The morphological tagging of the spoken corpus has been currently under the process of the testing.

2 Recording and Transcription

Each respondent is recorded in a formal (question-answer) and an informal (discussion between two or more people) interview. The text files are accompanied by notes specifying sex, age and education of respondents. Transcription of the tape recordings was performed in accordance with strict rules [2], which were created taking into consideration the specific nature of the spoken language. The rules are based on the transcription rules for the Prague Spoken Corpus [3], but they were slightly modified to respect the middle-Moravian interdialect more precisely. The main principle was to transcribe the speech as precisely as possible with a special emphasis on the articulation assimilation and simplified consonant clusters. However, even though the rules were always adhered to, transcription of some parts of the text depended on individual hearing impression (duration of vowels or pronunciation of frequent words).

3 Morphological Tagging

The Brno Spoken Corpus is grammatically annotated by a program which was created at the FI MU in Brno [4]. Each recognised word form is segmented into three formal parts: a stem base, an intersegment and an ending. The basic structure of this program consists of two parts: the dictionary of stem bases and the sets of patterns determining the only possible combinations of stem bases, intersegments and endings.

The spoken corpus must be annotated taking into account the milieu from which it was acquired. The spoken language in Brno includes three levels of vocabulary: literary Czech, colloquial Czech and the middle-Moravian interdialect. In the spoken corpus, there are words with non-literary endings, with changes in the stem base and with an individual pronunciation. These facts imply the necessity of adjustments to the program used for the tagging. Our target is to find the sets of these words which then could be annotated automatically as well.

3.1 Non-literary Endings

The set of endings in the morphological analyser, at first determined for written (and literary) texts, was extended by the non-literary endings for inflected words [5]. These endings were tagged by a special tag reflecting the occurrence in the respective vocabulary (we use `wC` for colloquial Czech and `wM` for the middle-Moravian interdialect). The following example shows the extension of the ending set `NS0216` for masculine adjectives in nominative singular.

```
=NS0216
 {ý,  k2gMnSc1d1}    # literary Czech
 {ej, k2gMnSc1d1wC}  # colloquial Czech
 {é,  k2gMnSc1d1wM}  # middle-Moravian interdialect
```

3.2 Changes in the Stem Base

The main sets of non-literary changes in the stem bases were selected. These changes do not cause any change in the flection of the given word. That is why we can only add the new stem base (with the change) to the existing pattern in the dictionary and the system of patterns can still remain untouched.

In the spoken corpus, there are very frequent groups of non-literary words (mainly non-inflected words) which are influenced by individual pronunciation. These forms are also included in the dictionary in all their occurrences with the corresponding lemma (and to the respective pattern), e.g. the word forms *prtože, prže, pže, prože, ptože* have the same (literary) lemma *protože/because* which denotes the conjunction. The basic sets of the non-literary changes in the stem bases follow:

1. prosthetic *v* - the phoneme is added to nearly every word beginning with the phoneme *o* (*okno - vokno/window*) except names and foreign words;
2. vocalic change *ou*>*ó*, *ý*>*ej*>*é* (*koupat - kópat/have a bath*, *týden - tejden - téden/week*) the most frequent change, specification of occurrence: e. g. superlative of adjectives and adverbs *nej*>*né* (*největší - névětší/the biggest*);
3. reduced vowels - e. g. one-syllabic infinitive with the word basis ending with a vowel (*dát - dat/to give*);
4. extended vowels - e. g. masculine singular, past participle ending with *l* (*bral - brál/he took*);
5. dialectic non-performing articulation dissimilation (*šťastný - ščastný/happy*);
6. articulation assimilation (*odstup - octup/distance*);
7. simplified consonant clusters (*elektrika - eletrika/electricity*) - the most frequent set of words.

One of the most important observations for further research is the frequency of particular types of changes. Such data are depicted in table 1.

Table 1. Frequency of non-literary changes in stem bases

non-literary change	**frequency (%)**
prosthetic *v*	0.76
vocalic change *ou*>*ó*, *ý*>*ej*>*é*	0.62
reduced vowels	0.31
extended vowels	0.01
dialectic non-performing articulation dissimilation	0.01
changes in the intersegment	12.60
changes in the ending	18.20

3.3 Changes in the Intersegment

We must pay attention to the fact that Czech words are not only a sum of stem bases and endings, there are intersegments between them, as well. This results in certain speech sound changes - alternation of a final speech sound of the stem. For this reasons, the morphological analyser was extended by a few new patterns that capture the changes in inflection. Typically, it is the case of very frequent verbs whose stems are dialectic or influenced by individual pronunciation, e.g. verbs *být/to be*, *chtít/to want* and *jít/to go*.

4 Experiment Results

We cannot expect that the automatic tagging of the spoken corpus would be 100 % successful (however, the rate for the written corpus is about 90 %). When we tested the analyser for the first time, 70 % words had the right tags at first sight. Nevertheless, the problems occur with words which have one word form (literary and non-literary) for two different meanings and which were automatically annotated as the literary ones by the program. As these words are very frequent, the real number of right tags is about 60 %. These words have had to be selected and the user must have a possibility to choose a right tag in accordance with the context, e.g. the word form *byt* is the noun *byt/flat* (literary) and the verb *být/to be* (dialectic). For these cases a special pre-program was created, which makes the right choice possible.

5 Conclusion

The morphological tagging of the spoken corpus has a nature of a process, i.e. the procedure is changed depending on the test results of the program using the language material. The problems which have appeared in the course of the testing are resolved by adjustments to the program on a continuous basis. The main requirement of the tagging is to reach a maximally automated analysis with a minimum need of human interference possible.

The spoken language, in contrast to the written one, represents a real use of natural speech. Unfortunately, the spoken corpora usually form a very small part of the large (written) corpora (about 1 % for Czech [3], less than 20 % for English in BNC [6]) because of the difficulty and high costs of recording and transcribing the natural speech. Nevertheless, the spoken corpora should be extended as well. Our target is to create the corpus of the spoken language from Moravia (about 1,000,000 word occurrences) in cooperation with the departments of the Czech language at the other Moravian universities (Palacký University in Olomouc and University of Ostrava).

References

[1] Rychlý, P.: GCQP – Multiplatform Graphical User Interface to the CQP Corpus Manager. In: Proceedings of the Ninth EURALEX International Congress, Institut für Maschinelle Sprachverarbeitung, Stuttgart (2000) 149–154

[2] Šonková, J.: Morfologie mluvené češtiny. PhD thesis, Filozofická fakulta Karlovy univerzity, Praha (1995)

[3] Čermák, F.: Language Corpora: The Czech Case. In: Proceedings of TSD 2001, Berlin, Springer-Verlag (2001) 21–30

[4] Sedláček, R., Smrž, P.: A New Czech Morphological Analyser `ajka`. In: Proceedings of TSD 2001, Berlin, Springer-Verlag (2001) 100–107

[5] Hlaváčková, D.: Korpus mluvené češtiny. Master's thesis, Filozofická fakulta Masarykovy univerzity, Brno (1998)

[6] Crowdy, S.: The BNC spoken corpus. In: Spoken English on computer: transcription, mark-up and application, Harlow, Longman (1995)