Abstract

The thesis aims at explicit description of Old Czech common nouns declension with regard to its application in a tool for automatic morphological analysis of (digitized) texts in Old Czech. This means that this description is intended to serve as a basis for automatic generation of word forms (jointly with their appropriate morphological information and lemma) which will then be used for assigning morphological categories (gender, number, case) and lemma to word forms occurring in Old Czech digitized texts. The thesis thus develops a base for the first step in transformation of text banks (which currently exist for the Old Czech period) into an Old Czech corpus offering more possibilities for linguistic research. The Old Czech period is defined as a period from the beginning of the 14th century (more precisely from the period when first coherent texts written in Czech appeared) approx. to the end of the 15th century. Nouns were chosen for this work, because they cover approx. 30% of texts in current Czech (which is the highest percentage from all parts of speech). Old Czech texts are taken into account only in a transcribed form (based on transcription rules used in the Old Czech Text Bank developed at the Institute of the Czech Language of the Academy of Sciences of the Czech Republic). On the one hand, the transcription greatly facilitates automatic morphological analysis, because it standardizes both characters and orthography, on the other hand, it is important to bear in mind that every transcription is by necessity an interpretation, and hence it is dependent, at least to some extent, on editor´s decisions.

Three sources of information were used for the Old Czech common nouns description – these are historical grammar books, Old Czech texts and dictionaries of Old Czech. The description of Old Czech common nouns declension in historical grammar books served as the starting point – the information provided was systematically checked and complemented using texts in the internal version of the Old Czech Text Bank. The version of this bank used for the majority of topics consisted of 7.6 mill. tokens. It was searched through by means of the Excel-based tool *Analýza tokenů (Analysis of tokens)* that enables a) to generate word forms on the grounds of a morphological basis (i.e. a part of a word common for all forms in a paradigm) and endings, and then to search for many word forms in texts at once, b) to display and search in lists of word forms filtered by sequences of characters at the end of word forms. Some texts in the range of approx. 3.2 mill. tokens of the internal version of the Old Czech Text Bank have not undergone a final editor check yet. Hence, in cases it was necessary to use some material from these texts, this material was always checked in copies of manuscripts or other sources of the original texts. The dictionaries of Old Czech accessible via web-interface *Vokabulář Webový (Web Vocabulary)* were used as the base for the list of Old Czech lemmata. However, as there is no dictionary covering the whole vocabulary of the Old Czech period, and the dictionaries differ in their methodology as well, the material extracted from them will require expansion and refinement.

The description of Old Czech common noun declension presented in this thesis consists of four parts.

In the first part, endings of declension types (which are historically stem-based) are described both in a text and in a tabular form. The historical origin and the positive text evidence of each ending are displayed in these tables as well. Each declension type consists of different number of declension patterns. A declension pattern is defined as a distinct set of endings which is used for a particular set of

common nouns (e.g., nouns for persons, or nouns with stem final velar consonant) to create all word forms. Overall, 96 declension patterns of 22 declension types were described (the masculine o-stems, neuter ьjo-stems and feminine a-stems being the most numerous declension types).

The second part describes alternations, i.e. changes in a morphological base that are impossible to apply in a form of a general rule because they do not apply to all lemmata with a given form (cf. *pes-Ø* [dog-NOM.SG] – *ps-a* [dog-GEN.SG], but *les-Ø* [forest-NOM.SG] – *les-a* [forest-GEN.SG]; *kráv-a* [cow-NOM.SG] – *krav-ám* [cow-DAT.PL], but *krás-a* [beauty-NOM.SG] – *krás-ám* [beauty-DAT.PL]), or their application in a form of a rule would be too complicated (cf. *hvězd-a* [star-NOM.SG] – *hvězd-Ø* [star-GEN.PL], *otázk-a* [question-NOM.SG] – *otázek-Ø* [question-GEN.PL], *šacht-a* [shaft-NOM.SG] – *šacht-Ø*/*šachet-Ø* [shaft-GEN.PL]). Alternations are described in a text and for each type of alternation a special label is used in the list of common noun lemmata (the fourth part of the thesis) as a signal that the given alternation applies to the given lemma. Overall, approx. 120 types of alternations were described (the alternation caused by yer vocalization and subsequent analogical development is applied for the highest number of lemmata).

Sound changes are described in the third part. They were defined as formal changes of word forms which can be described by a general rule and they comprise both i) changes connected with language development in the Old Czech period (e.g., *viera* – *víra* [faith], *bóh* – *buoh* [god]), forms presumed for year 1300 being considered as base forms, and ii) changes occurring as a result of connecting a morphological base and an ending, some of these changes being a matter of orthography only (e.g., the morphological base *vlk* [wolf] connected with the ending *-i* [NOM.PL] results in the form *vlci*, or the base *líň* [tench] connected with the ending *-em* [INS.SG] results in the form *líněm*). Sound changes are described both in a text and in a form of rules specifying the changes of letters (taking into account context of the change, if necessary). Overall, almost 100 rules for sound changes were described.

The list of common nouns lemmata represents the fourth part of the thesis. Here, lemmata are assigned to declension patterns and to alternation if there is any in the paradigm of the given lemma. The first version of this list was created by means of automatic extraction of common nouns lemmata from dictionaries of Old Czech and it was then sorted and extended manually. It consists of approx. 29,000 lemmata. In connection with other parts, this list will serve as a base for automatic generation of word forms: morphological bases will be extracted from lemmata and according to information on declension pattern, they will be combined with endings (with regard to eventual alternation), a correct connection of a base and an ending will be assured by sound changes and application of these changes will provide all other possible forms of the given word form as well.

As a complement, the thesis contains also a list of exceptions whose systematic treatment would be pointless, as the exceptions are irregular and rare word forms.

The main asset of the presented approach to the development of a tool for tagging and lemmatization is a creation of a complex description of formal morphology for the given period which makes it possible to use detailed linguistic information (declension type, sound/orthography changes) in the automatic morphological analysis. The price to be paid for such an approach, however, lies in its time consuming character and its direct dependency on sources upon which it is based. Therefore the presented description is meant as a base which will necessarily have to be updated and completed in connection with the development of sources used.

At a more general level, the thesis tests the presented approach as a whole – if it becomes the base for a reliable tool for automatic morphological analysis of Old Czech common nouns, the similar/same approach will be used for other parts of speech as well in the future.