Describing Old Czech Declension Patterns for Automatic Text Analysis

Pavlína Jínová, *Charles University, Faculty of Arts (Czech Republic)*
jinova@ff.cuni.cz

Boris Lehečka, *Institute of the Czech Language, Academy of Sciences of the Czech Republic (Czech Republic)*
boris@daliboris.cz

Karel Oliva, *Institute of the Czech Language, Academy of Sciences of the Czech Republic (Czech Republic)*
kareloliva@gmail.com

Received: August 4, 2014.
Accepted: November 17, 2014.

ABSTRACT
This paper focuses on describing declension patterns in Old Czech (West Slavonic language, stadium between approx. 1300 and 1500) which is one part of the base serving the automatic annotation of the Old-Czech Text Bank. We introduce our approach to automatic morphological analysis, its principles (historical justifiability, constant regard to system of the language and systematic account of phonological changes) and subparts (a dictionary, a description of patterns, a list of stem changes accompanying declension, rules for sound changes, and the list of exceptions). We also illustrate the process of searching for declension patterns of two feminine declensions in complicated mutual interaction – i-stems and ja-stems. We came to the conclusions that borrowed endings documented in available resources (forms like *zem-ech* (LOC.PL, combination of *ja*-stem noun *zem-ě* (a land) with *i*-stem ending *-ech*) are not exceptions; they belong to the system of language and the base for automatic analysis should treat them as such.

Key words: Old Czech morphology, declension patterns, automatic text analysis, i-stems, ja-stems.

1. Introduction

The aim of this paper is twofold: first, we briefly present a linguistically oriented approach to Old Czech (a West Slavonic language) morphology, which serves for the automatic analysis of Old Czech texts (approx. from the beginning of 14th century to 1500); we describe the main principles of this approach and all its subparts (Section 2). The main part of this paper is then devoted to the description of declension patterns, which is one of the subparts of the whole approach. We illustrate the process of searching for sufficient descriptions of a declension pattern by using examples of two Old Czech feminine declensions in complicated mutual interaction, namely i-stems and ja-stems (Section 3).

2. A linguistically oriented approach to Old Czech morphology

2.1. Motivation

Texts written at older stages of modern languages create a challenge for automatic processing (such as morphological analysis, tagging, parsing, etcetera) due to the (natural) lack of large-scale resources such as electronic lexica, pre-annotated training data, etc. These problems are solved in various ways. Most often tools developed for modern languages are used for analysis and results are afterward supplemented by some manual annotation.

Good examples of this approach from recent years are The Quranic Arabic Corpus (Dukes and Habash, 2010), The Russian Diachronic Online Corpus (Meyer, 2011), or an attempt to annotate Old Czech texts using tools for modern Czech (Hana et al., 2012).

As opposed to these approaches, we propose an approach based on a description oriented on capturing the equilibrium of stability and change (and resulting diversity) of a language system over a long period. We dare to claim that the very nature of diachronic investigation of language is somehow different from the investigation of its modern stage – there are no native speakers whose intuition could be used as an arbiter in case of analysis of questionable cases and moreover, there is only a limited amount of data (texts) any research can rely on. These facts make each documented text (and each word-form contained) more important than in modern stages of language (in this respect we completely agree with Meyer (2011) that for diachronic corpora, "expectations on the precision of markup are generally higher than for large contemporary corpora").

From our point of view, there is therefore a need to capture not only the core of language which can be analyzed by tools developed for modern stages (cf. 74% performance of Hana's (Hana et al., 2012) tagger using tools developed for Modern Czech), but also exceptional cases, minor forms etc., which are in fact more substantial for linguistic research as they can reveal important facts about the nature of historical stages and development of the particular language. We thus work with Old Czech vocabulary instead of a Modern one, we define forms of stems according to linguistic description (Gebauer, 1960 and 1963) and constantly work with sound changes and stem alterations (see below). The general approach is applicable to all parts of speech, even though in practice we have worked with common nouns only so far.

2.2. Data

Old Czech is a morphologically rich West Slavonic language. With regard to the morphology of nouns, which constitutes the core of our current research, Old Czech has seven cases and three genders (masculine, feminine, neuter) as does Modern Czech, while it differs from Modern Czech in the fact that three numbers (singular, dual, plural) are part of the grammatical system. Declension types are stem-based, 22 different declension types are traditionally postulated for this stage of Czech language (Gebauer, 1960, Vážný, 1970, Lamprecht et al., 1986).

Old Czech manuscripts and incunabula are successively being made accessible in a searchable electronic format by the Institute of the Czech Language. Texts are transcribed using modern Czech orthography (cf. Lehečka & Voleková, 2010) and are included in the Old-Czech Text Bank, in the section of web resources for Old Czech (http://vokabular.ujc.cas.cz). So far, 150 Old Czech documents (approx. 3 million Czech word tokens) have been processed and included into (the publicly available part of) the Bank, which however in its ultimate stage aims at incorporating all documents surviving from the beginning of written Czech to the end of the 15th century (it could be slightly more than 1230 documents). An internal version of the Text Bank, used for research into i-stem and ja-stem declension presented below contains also texts which have not undergone final correctness checking and consists of 231 documents of the Old Czech time period (approx. 6 million of Czech word tokens). The Text Bank contains a wide variety of different genres, but it is not balanced with respect to the periods in which the texts originate.

2.3. Parts of analysis

For an appropriate morphological analysis of common nouns occurring in Old Czech documents, we decided first to create procedures for generating all those forms which can be fairly hypothesized given the current evidence and second to detect these forms in documents and assign the morphological information (gender, case and number), lemma and hyperlemma to these occurrences (as a hyperlemma we use the form appropriate for the year 1300). We use one basic principle for generating forms of each lemma: they should be to the maximum extent historically justifiable, that is: preferably attested to in texts of the Text Bank (as available at the current stage of its development) or in other historical texts, or mentioned in traditional linguistic handbooks and/or systematic w.r.t. attested forms of "similar" lemmata (we use limited resources as a starting point, but intend to cover all texts from the Old Czech period). A systematic part of our work is the analysis of patterns of sound changes, because we believe that this should reveal some linguistically interesting phenomena such as combination of changes from different periods in one document or even in one word. This approach brings as its by-product a systematic description of Old Czech morphology as well as reducing the complexity of morphological description.

As basis for automatic analysis, we use:

1) The electronic dictionary was developed by the Institute of Czech language and it covers most and will cover (by 2018) all documented Old Czech vocabulary. Each noun item of the dictionary contains information about gender and ending in GEN.SG (crucial clue for determining the declension type). To date, the Electronic dictionary covers more than 69.5% of the vocabulary, the remaining part is captured by other digitized dictionaries (for details see http://vokabular.ujc.cas.cz/zdroje.aspx). For the majority of declension types, the dictionary contains all the necessary information for assignment of each lemma to appropriate declension type. Some types of declension need, however, more detailed types of information (e.g. the distinction person/animal/non-living object is relevant for certain classes of masculine nouns (o-stems)).

2) A detailed description of endings for each type of declension based on available linguistic descriptions and complemented by results of manual text analysis. For each type of declension, we created i) a basic set of endings which can be combined with all stems belonging to a specific type of declension, ii) a set of endings with some restriction (e.g. ending *-u* in GEN.SG of o-stem masculine is combinable only with nouns for animals and non-living objects, not with nouns denoting persons; for ja-stem feminine nouns, GEN.PL without ending is documented only with stems ended by *j*, *l*, *m*, *n*, *ž*, *š*, *c* and so on).

3) The rules for stem alteration. Some stems in Czech change their form according to phonological properties of endings or of the whole form (e.g. number of syllables). There are two types of stem alteration rules: i) regular changes at the end of stem, which can be well described by simple rules (e.g. each stem-final in *h*, *ch*, *k*, *r* combined with endings *-i*, *-ě*, *-ie* or their later forms changes respectively to *z*, *š*, *c*, *ř* (e.g. *břuch-o* (NOM.SG, a belly) > *břuš-ě* (LOC.SG)), ii) alterations inside the stem which are more complicated and rules for them have to be found only after text analysis (e.g. some forms of GEN.PL without ending

have stems that undergo no change as compared to NOM.SG (e.g. *měst-o* (NOM.SG, a city) > *měst* (GEN.PL)), some of them display only a form with *e*-alteration (e.g. *stehn-o* (NOM. SG, a thigh) > *stehen* (GEN.PL)) and some stems show both forms (e.g. *křídl-o* (NOM.SG, a wing*)* > *křídl* and *křídel* (both GEN.PL)).

4) Sound changes. Old Czech texts cover well over two centuries of language history – numerous sound changes applied in stems and also in endings during this period, some of them being context free (e.g. change *ie* > *í*, *ie* > *é*, normal in 15th century, however, occurring rarely from as early as the end of 13th century), some of them are context sensitive (e.g. change *ě* > *e* after sibilants and liquids). We include these changes into procedures of derivation of stems (e.g. stem of lemma *střiebro* (silver) can have forms *střiebr-*, *stříbr-*, *střébr-*) and also derivation of endings (e.g. ending for LOC.PL in several declensions *-iech* can have also forms *-éch* and *-ích*).

5) List of exceptions. For the majority of declension types, some exceptional forms are documented: typically some solitary combinations of stem from one declension type with ending from another type (e.g. o-stem neuter noun *čísl-o* (NOM.SG, a number) has the documented form *čísl-u* in GEN.SG (Gebauer, 1960: 136), though *-u* is u-stem masculine ending, the regular form is in this case *čísl-a* (GEN.SG)). We collect these exceptional forms and make them part of our analysis.

3. Describing Old Czech declension patterns for i-stems and ja-stems

3.1. Introduction

After having introduced the approach as a whole, we now illustrate the process of searching for sufficient declension patterns (part 2 of the whole analysis, see above) using two groups of Old Czech feminine nouns, i-stems and ja-stems, as an example.

In Old Czech, i-stems are for example words as *věc* (a thing), *kost* (a bone), *bolest* (pain), *dan* (a tax), *tiesn* (distress), or *sól* (salt), ja-stems words as *tvrzě* (a stronghold), *země* (a land), *chvíle* (a while), *péčě* (care), *róžě* (a rose), or *uličě* (a street). Simplified paradigms of ja-stems and i-stems are given in Table 1. It shows proper endings for both declensions (according to Gebauer, 1960: 201–202, 342–343), but it does not include either form of endings after sound changes or endings for dual number, because this information is not relevant here.

	ja-stems	i-stems
NOM.SG	*-ě*	*Ø*
GEN.SG	*-ě*	*-i*
DAT.SG	*-i*	*-i*
ACC.SG	*-u*	*Ø*
VOC.SG	*-e*	*-i*
LOC.SG	*-i*	*-i*
INS.SG	*-ú*	*-ú*
NOM.PL	*-ě*	*-i*
GEN.PL	*Ø, -í*	*-í*
DAT.PL	*-iem*	*-em*

ACC.PL	*-ě*	*-i*
VOC.PL	*-ě*	*-i*
LOC.PL	*-iech*	*-ech*
INS.PL	*-ěmi*	*-mi*

Table 1. Proper endings of i-stems and ja-stems in Old Czech.

I-stems and ja-stems have always been in complicated mutual interaction. In modern Czech, this interaction is most often viewed as a gradual development of a new declension pattern, i.e. a pattern with NOM, ACC and VOC.SG from i-stems and the rest of paradigms from ja-stems (a detailed account of this question for Old and Middle Czech is given in Vajdlová, 2012 and 2013). At the stage in question, however, the interaction of i-stems and ja-stems is not limited to the development of a new declension pattern – it is substantially broader. Gebauer in his most reliable and extensive historical grammar of Czech (Gebauer, 1960: 202–214, 343–349) demonstrates that in the stage of language in question, i-stems can borrow endings from ja-stems and ja-stems from i-stems in almost all cases where these endings differ, i.e. in NOM, GEN and ACC.SG, and in NOM, ACC, DAT, LOC and INS. PL (GEN.PL with zero ending for i-stems was not documented as well as borrowings in VOC.SG and VOC.PL, we therefore do not include these cases into our analysis below). So we encounter forms as *tvrzech* or *věcemi* in Old Czech texts, which we would not expect according to the proper forms of i-stems and ja-stems: *tvrz-ech* is LOC.PL of ja-stem *tvrz-ě* (a stronghold) with i-stem ending *-ech*, the regular form would be in this case *tvrz-iech*; *věc-ěmi* is INS.PL of i-stem *věc* (a thing) with ja-stem ending *-ěmi*, the regular form would be in this case *věc-mi*.

However, documentation in available resources (Gebauer, 1960, Vážný, 1970, Lamprecht et al., 1986) is insufficient for the purposes of automatic morphological analysis, since it is not detailed enough. In other words, we simply do not know if these borrowings are only exceptions or if they represent some common phenomenon in the Old Czech declension system. This difference has serious consequences for automatic text analysis: if these forms were exceptions, they would be part of the list of exceptions only, if they, on the other hand, were some part of the system, they should become a part of the paradigm, i.e. a base to form generation. In the latter case we would like also to know if the borrowed endings should be applied to all i-stem and ja-stem lemmata or if there is some regularity of their spread – e.g. some lemmata or groups of lemmata allow for them and others do not.

3.2. Procedure

Gebauer (1960: 212 and 342) suggests that the clue for borrowing lies in final consonants of the stem: ja-stems with final stem sibilants (e.g. nouns as *ulic-ě* (a street)) often have i-stem ending and i-stems with *n* as final stem consonant (e.g. nouns as *tiesn* (distress)) often have ja-stem ending. Therefore we decided to explore whole situation using this criterion systematically.

First, we looked up all possible final stem consonants for ja-stems and i-stems. Then, we systematically searched for all combinations of these consonants and endings which concern borrowing between declensions – we searched for borrowed endings in GEN.SG, NOM.PL, ACC.PL, DAT.PL, LOC.PL and INS.PL (NOM and ACC.SG have also different endings for both declensions – see Table 1, but they were omitted, because the situation in these cases is sufficiently captured by the dictionary). Forms with these endings for each stem final consonant were searched via a database of all tokens from the internal version of the Old-Czech Text Bank, for example all tokens ended by *-zech* were found in this database (both i-stems and ja-stems have lemmata with stems ended by *z*, *-ech* is i-stem ending of LOC.PL, which is a case where borrowing can occur) and we selected manually all ja-stem forms (as *tvrzech*) from them. All forms found were then verified in texts of the Old-Czech Text Bank. For all of them, we had to check the context (e.g. for ja-stems, GEN.SG with borrowed ending (*obc-i*, NOM.SG *obc-ě* (a community)) is ambiguous with DAT.SG and LOC.SG), some of them had to be verified by manual analysis using original texts, because digitized texts have not yet undergone final checks by experts in transcription. To forms found in Old-Czech Text Bank, forms cited in Gebauer's work were added to obtain a more extensive picture of the whole situation. Finally, the tables with verified documented forms were created distinguishing different cases and different final stem consonants.

3.3. Results

There are approx. 1 700 different ja-stem lemmata and 2 000 different i-stem lemmata in Old Czech dictionaries (digitized versions of these dictionaries are available through http://vokabular.ujc.cas.cz), they do not differ in the repertoire of final stem consonants – both groups have stems ended by an identical types of sibilants (*c*, *č*, *s*, *š*, *z*, *ž*, *ř*), dentals (*d*, *t*, *n*), liquids (*l*, *j*) and labials (*p*, *b*, *m*, *v*). (These types are established here according to the tradition of Czech historical grammars.)

Ja-stems and i-stems documented with borrowed endings are included in Table 2 and Table 3. The first column in both tables shows the final stem consonant, the second column displays the approximate number of all lemmata with this final stem consonant. The rest of the columns are then devoted to forms found for each case where the borrowing of endings can occur, i.e. for GEN.SG, NOM.PL, ACC.PL, DAT.PL, LOC.PL and INS.PL. All forms found are included. Table 2 illustrates the situation for ja-stem lemmata with i-stem endings, Table 3 has the opposite situation – i-stem lemmata with ja-stem endings.

ja-stem lemma with i-stem ending							
final stem C	number of lemmata	case.number and ending					
		GEN.SG	NOM.PL	ACC.PL	DAT.PL	LOC.PL	INS.PL
		-i	*-i*	*-i*	*-em*	*-ech*	*-mi*
c	1100	*bukvici* *obci* *pravici* *róznici* *studnici* *ulici* *unci*	*práci* *holubici*		*bláznicem* *krticem* *lavicem* *obcem* *opicem* *ovcem* *plicem* *pracem* *Vánocem* *vdovicem*	*plecech* *plicech* *pracech* *stolicech* *ulicech* *ovcech* *vinicech*	*lžicmi* *svěcmi* *sviecmi*

č	27	*húšti* *katrči* *púšči* *šíři*				*péčech*	*kleštmi* *péčmi*
z	23	*mezi* *núzi* *rzi* *tvrzi* *žiezi*			*mezem* *nesnázem* *tvrzem*	*tvrzech*	*tvrzmi*
ž	41	*blíži* *kuoži* *níži stráži*				*strážech* *věžech*	*kožmi* *věžmi*
s	2						
š	41	*duši* *mši* *peleši* *skrýši*			*dušem* *mšem*	*dušech* *mšech* *pelešech*	*dušmi*
ř	28	*číři* *šíři* *večeři* *záři* *zoři*			*přem*	*přech*	*zářmi*
l	159	*dáli* *chvíli kdúli* *lavenduli* *míli* *posteli*		*dáli*	*košilem* *nedělem*	*maštalech* *mílech* *nedělech* *úlehlech*	*nedělmi*
j	58	*veřeji*			*šlépějem*	*veřejech*	*lilijmi*
d	2					*paždech*	*paždmi*
t	1						
n	190	*jeskyni* *škráni* *vuoni*					
p	12					*náspech*	
b	11	*ohbi*					
v	1						
m	4	*krmi* *zemi*				*zemech*	

Table 2. Ja-stem lemmata with i-stem ending according to the internal Old-Czech Text Bank (versions 10.12.2013, 21.02.2014 and 02.05.2014) and Gebauer (1960: 202–227).

As Table 2 shows, ja-stem lemmata with i-stem endings are well documented in the case of stems ended by sibilants and *l*. Other types of consonants seem not to support borrowing, but on the other hand they do not seem to make it impossible. We have to take into account also the number and frequency of lemmata for each type of consonants. For example, there is only one lemma with a stem ending in *t* and one ending in *v*, and moreover, both of them are quite rare, so we cannot expect their occurrence with borrowed ending at all. A type attracting certain attention is ja-stem stems ending in *n* – they are quite numerous but they are documented only rarely and only in GEN.SG with borrowed endings. This situation can be in our opinion caused by the infrequent occurrence of the majority of these lemmata in texts. The final observation we could make on the basis of Table 2 is that the occurrence of borrowed endings in NOM and ACC.PL is quite rare.

i-stem lemma with ja-stem ending							
final stem C	number of lemmata	case.number and ending					
		GEN.SG	NOM.PL	ACC.PL	DAT.PL	LOC.PL	INS.PL
		-ě/-e	*-ě/e*	*-ě/e*	*-iem/-ím*	*-iech/-ích*	*-ěmi/-emi*
c	21	*pece*	*moce věce*		*věciem*	*nemociech* *nociech* *věciech*	*mocemi* *pomocemi* *věcemi*
č	54	*obruče* *vodoteče* *zeměžluče* *žlučě*		*pavlačě*	*řečiem*	*řečiech*	*loučemi* *řečemi*
z	6						
ž	8	*lžě*					
s	5				*vsím*	*vsiech*	
š	12	*rozkošě*				*myšiech*	
ř	14	*dormitoře* *lektvaře* *tváře* *zvěřě*	*tváře*	*tvářě*	*tvářiem* *zvěřiem*	*lektvařích* *tvářích*	*tvářemi*
l	33	*kúpěle* *ocele*				*holích*	*húslemi* *ratoraslemi*
j	1						
d	31	*čeledě* *zpovědě*	*kádě*	*odpovědě* *piedě*		*přípovědiech*	*zděmi*
t	1670	*rtutě* *sietě* *smrtě*		*plstě*	*čělistiem* *poutím*	*jěšutiech* *lestiech* *vlastiech* *žalostiech*	*bolestěmi* *hořkostěmi*
n	91	*básně* *bázně* *dásně* *dlaně* *holeně kázně* *piesně* *pláně* *přízně* *sieně* *tiesně žiezně* *žně*	*daně dásně* *holeně* *piesně* *tiesně*	*daně* *dásně* *holeně* *nepřiezně* *piesně* *písně* *sieně* *trýzně* *žně*	*básniem* *daniem* *dásniem* *písniem*	*básniech* *daniech* *dásniech* *neprázdniech* *piesniech* *sieniech*	*bázněmi* *braněmi* *daněmi* *dásněmi* *dlaněmi* *piesněmi* *rovněmi* *saněmi*
p	4						
b	14	*korábě zlobě*					
v	11						
m	2						

Table 3. I-stem lemmata with ja-stem ending according to the internal Old-Czech Text Bank (versions 10.12.2013, 21.02.2014 and 02.05.2014) and Gebauer (1960: 342–401).

Table 3 differs from Table 2 in some respects. Forms with borrowed endings are the most numerous for i-stem stems ending in *n*. This exceptional position of stems ending in *n* is well observable in comparison with other stems ending in dentals – for stems ending in *d* and *t*, there are on the one hand forms documented with borrowed endings, but, on the other, especially for *t*, their number is quite small in comparison with number of lemmata. Less i-stem lemmata than ja-stem lemmata have stems ending in sibilants, but borrowed endings are here quite well documented, too. As well as for ja-stems with i-stem ending, stems ending in labials are documented with borrowed ending only rarely. This could however be caused again by the small number of lemmata and their low frequency of occurrence in Old Czech texts. A similar situation can be observed for stems ending in *j* – there is only one lemma with low frequency of occurrence, so we did not find any case of borrowing. As well as for ja-stems with i-stem endings, i-stem stems ending in *l* are quite well documented with ja-stem endings. To sum up, we can state that no type of consonants seems to prevent borrowing, as we observed also for ja-stems with i-stem endings. NOM.PL and ACC.PL are documented with borrowed endings less frequently than other cases but not as rarely as we saw for ja-stems with i-stem endings. Especially for i-stem stems ending in *n* these forms seem to be quite normal with a borrowed ending.

To conclude, we can state that we confirmed the observations made by Gebauer that borrowed endings are well documented especially for ja-stems with stems ending in sibilants and for i-stems with stems ending in *n*. On the other hand, however, we saw that any type of consonant does not strictly prevent borrowing. We also observed that borrowed endings are quite rare for NOM and ACC.PL concerning ja-stems with i-stem endings, but that for i-stems with ja-stem endings the situation differs.

3.4. The Solution for automatic morphological analysis

Taking into account these findings we can formulate three ways of finding a solution for automatic text analysis. The first way is to connect borrowed endings with documented forms only – we can include borrowed forms only for lemmata with which they were documented. Although this solution is the best one from the perspective of observation adequacy, it is unsatisfactory due to the fact that it is based on a restricted collection of texts – every new text added to this collection could change the situation.

The second way is to include borrowed endings only for lemmata with stems ending in some type of consonants (e.g. for ja-stem stems ending in sibilants or i-stem stems ending in *n*). Although our analysis revealed that for some consonants (i.e. for labials) borrowings are documented rather rarely, we cannot claim that these forms do not occur in texts. In other words, we cannot be sure that we can omit some types of lemmata and still capture all forms in texts.

The third solution in the most complex one and in our opinion also in best fit with our findings – we should include borrowed endings with all lemmata of both declensions, because we cannot be sure for which lemmata/types of stems borrowing is truly ruled out. This solution brings along the problem of the arising of ambiguous forms; in our opinion, however, this problem can be solved more easily than the problem of non-analyzed forms which would arise if borrowed endings were not included in paradigms of ja-stems and i-stems.

4. Conclusion

This paper deals with describing Old Czech (a West Slavonic language, stadium between approx. 1300 and 1500) declension patterns for automatic morphological analysis of Old Czech texts. We first described the whole approach to this analysis, its principles and subparts. We argued that preparing the base for automatic analysis concerning all diversity of language system, respecting forms documented in texts and constantly regarding system and sound changes could serve automatic analysis better than mere transformation of tools used for analysis of the modern stage of the language. As base for automatic analysis, five subparts were introduced: the dictionary for the language stage in question, the detailed description of endings based on available grammatical description and complemented by results of manual text analysis, the detailed description of stem changes accompanying declension, the rules for sound changes and the list of exceptions.

In the second part of the contribution, we illustrated the process of searching for declension pattern using the example of two feminine declensions with richly documented borrowing of endings – i-stems and ja-stems. The problem to be solved was whether the documented forms with borrowed endings are only exceptions and should be accordingly included in list of exceptions only, or whether they represent some systematic phenomena in Old Czech and should be treated as integral parts of i-stem and ja-stem paradigms. As a clue for our analysis, types of final stem consonants were adopted. We searched for all possible final stem consonants for i-stems and ja-stems, combined them with borrowed endings, searched these combinations in database of tokens from Old Czech texts and verified forms found in Old Czech texts. Contrary to our expectations, forms with borrowed endings do not seem to follow any abstract pattern – no type of final stem consonant prevents borrowing completely, although some types of consonants are attested with borrowed endings (especially ja-stem stems ending in sibilants and i-stem stems ending in *n*) more often than others (stems ending in labials are for both declensions documented with borrowed ending quite rarely). Taking into account these findings, we decided to include borrowed endings into ja-stem and i-stem paradigms systematically to be the best solution for automatic analysis, although it brings along increased ambiguity, because otherwise we could omit some forms in newly added texts.

Considering the whole process, we are aware of its time-consuming character, but we believe that due to the fine-granularity of the final analysis, it should reveal some interesting facts about Old Czech and bring a detailed description of Old Czech morphology as its by-product. Moreover, as our work is based on a lexicographically processed vocabulary, description of sound changes and detailed analysis of declension types, each non-detected form in our approach calls attention to three situations important for research of Old Czech: i.e. a)a new lemma, b) a new form different from those captured by traditional description, c) the mistakes made by editors of text.

Acknowledgement

This paper is an output of the project »Automatic Morphological Analysis of Old Czech Common Nouns« realized at the Faculty of Arts of Charles University in Prague and supported by the Specific University Research resources in 2014. The third author was supported by grant LM2010013 LINDAT/CLARIN awarded by the Ministry of Education, Youth and Sports of the Czech Republic.

REFERENCES

Dukes, K., & Habash, N. (2010). *Morphological Annotation of Quranic Arabic*. Paper presented at the 7th International Conference on Language Resources and Evaluation, La Valetta, Malta. Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/.

Gebauer, J. (1960). *Historická mluvnice jazyka českého. Díl III. Tvarosloví. I. Skloňování*. Praha: Nakladatelství Československé akademie věd.

Gebauer, Jan (1963). *Historická mluvnice jazyka českého. Díl I. Hláskosloví*. Praha: Nakladatelství Československé akademie věd.

Hana, J., Lehečka, B., Feldman, A., Černá, A., & Oliva, K. (2012). *Building a Corpus of Old Czech*. Paper presented at the Adaptation of Language Resources and Tools for Processing Cultural Heritage Objects Workshop associated with the 8th International Conference on Language Resources and Evaluation, Istanbul, Turkey. Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/workshops/13.ProceedingsCultHeritage.pdf.

Lamprecht, A., Šlosar, D., & Bauer, J. (1986). *Historická mluvnice češtiny*. Praha: SPN.

Lehečka, B., & Voleková, K. (2010). (Polo)automatická počítačová transkripce. In: M. Čornejová, L. Rychnovská & J. Zemanová (Eds.), *Dějiny českého pravopisu (do r. 1902)*. Brno: Host, 466–478.

Meyer, R. (2011). Old wine in new wineskins? Tagging Old Russian via annotation projection from modern translations. *Russian Linguistics,* 35 (2), 267–281.

Staročeská textová banka [on-line] [Old-Czech Text Bank]. Version 26. 06. 2014 and previous [cited 30. 06. 2014]. Oddělení vývoje jazyka Ústavu pro jazyk český AV ČR, v. v. i. Available at *http://vokabular.ujc.cas.cz/banka.aspx*.

Vajdlová, M. (2012). O formování nekmenového deklinačního typu *píseň* (se zaměřením na období staro- a středněčeské). In: S. Čmejrková, J. Hoffmanová & J. Klímová (Eds.), *Čeština v pohledu synchronním a diachronním: stoleté kořeny Ústavu pro jazyk český*. Praha: Karolinum, 184–189.

Vajdlová, M. (2013). *Ja*-kmenová feminina s alternativním zakončením nom. sg. *-ě/0* ve vztahu k deklinačnímu typu *píseň* (pohled vývojový). *Lingvistika Praha*, 2013. Retrieved from *http://lingvistikapraha.ff.cuni.cz/node/180*.

Vážný, V. (1970). *Historická mluvnice česká. Díl 2. Tvarosloví. 1. část*, *Skloňování*. Praha: SPN.

Vokabulář webový [on-line] [Web Dictionary]. Version 0.9.1. [cited 30. 06. 2014]. Oddělení vývoje jazyka Ústavu pro jazyk český AV ČR, v. v. i. Available at *http://vokabular.ujc.cas.cz*.