

Building a Corpus of Old Czech

Jirka Hana,¹ Boris Lehečka,² Anna Feldman,³ Alena Černá,² Karel Oliva²

¹Charles University, MFF, Prague, Czech Republic

²The Academy of Sciences of the Czech Republic, Institute of the Czech Language, Prague, Czech Republic

³Montclair State University, Montclair, NJ, USA

Abstract

In this paper we describe our efforts to build a corpus of Old Czech. We report on tools, resources and methodologies used during the corpus development as well as discuss the corpus sources and structure, the tagset used, the approach to lemmatization, morphological analysis and tagging. Due to practical restrictions we adapt resources and tools developed for Modern Czech. However, some of the described challenges, such as the non-standardized spelling in early Czech and the form and lemma variability due to language change during the covered time-span, are unique and never arise when building synchronic corpora of Modern Czech.

Keywords: Old Czech; Corpus; Morphology

1. Introduction

This paper describes a corpus of Old Czech and the tools, resources and methodologies used during its development. The practical restrictions (no native speakers, limited amount of available texts and lexicons, limited funding) preclude the traditional resource-intensive approach used in the creation of corpora for large modern languages. However, many high-quality tools, resources and guidelines exist for Modern Czech, which is in many aspects similar to Old Czech despite 500 years of development. This means that most tools, etc. do not need to be developed from scratch, but instead can be based on tools for Modern Czech.

Our paper is structured as follows. We outline the relevant aspects of the Czech language and compare its Modern and Old forms (§2.). We describe the sources and basic attributes of the corpus (§3.); lemmas and tagset used in annotation (§4.); semi-manual lemmatization (§5.); and finally, resource light morphological analysis and tagging based on Modern Czech and its more resource-intensive improvement (§6.).

2. Czech

Czech is a West Slavic language with significant influences from German, Latin and (in modern times) English. It is a fusional (inflective) language with rich morphology, a high degree of homonymy of endings and so-called free word-order.

2.1. Old Czech

As a separate language, Czech forms at the end of the 10th century AD. However, the oldest surviving written documents date to the early 1200's. The term Old Czech (OC) usually refers to the language as used roughly between 1150 and 1500. It is followed by Humanistic Czech (1500-1650), Baroque Czech (1650-1780) and then Czech of the so-called National Revival. Old Czech was significantly influenced by Old Church Slavonic, Latin and German.

2.2. Modern Czech

Modern Czech (MC) is spoken by roughly 10 million speakers, mostly in the Czech Republic. For a more de-

tailed discussion, see for example (Naughton, 2005; Short, 1993; Janda and Townsend, 2002; Karlík et al., 1996). For historical reasons, there are two variants of Czech: Official (Literary, Standard) Czech and Common (Colloquial) Czech. The official variant is based on the 19th-century resurrection of the 16th-century Czech. The two variants are influencing each other, resulting in a significant amount of irregularity, especially in morphology. The Czech writing system is mostly phonological.

2.3. Differences

Old Czech differs from Modern Czech in many aspects, including orthography, phonology, morphology and syntax. Some of the changes occurred during the period of Old Czech. Providing a systematic description of differences between Old and Modern Czech is beyond the scope of this paper. Therefore, we just briefly mention a few illustrative examples. For a more detailed description see (Vážný, 1964; Dostál, 1967; Mann, 1977).

2.3.1. Phonology and Spelling

Examples of some of the more regular sound changes between OC and MC can be found in Table 1. Moreover, the difference in the pronunciation of *y* and *i* is lost, with *y* being pronounced as *i* (however, the spelling still in most cases preserves the original distinction). In addition to these linguistic changes, the orthography develops as well; for more details, see (Křístek, 1978; Kučera, 1998).

2.3.2. Nominal Morphology

The nouns of OC have three genders: feminine, masculine, and neuter. In declension they distinguish three numbers: singular, plural, and dual, and seven cases: nominative, genitive, dative, accusative, vocative, locative and instrumental. Vocative is distinct only for some nouns and only in singular.

During the Old Czech period, the declension system moves from a noun-to-paradigm assignment based on the stems to an assignment based on gender. The dual number is replaced by plural, e.g., OC: *s jedinýma dvěma děvečkama* vs. MC: *s jedinými dvěma děvečkami* 'with the only two maids'. In MC, the dual number survives only in declension of a few words, such as the paired names of parts of

change during OC	later change	example	
<i>ú</i> > <i>ou</i>		<i>múka</i> > <i>mouka</i>	‘flour’
<i>’ú</i> > <i>’í</i>		<i>kl’úč</i> > <i>klíč</i>	‘key’
<i>sě</i> > <i>se</i>		<i>sěno</i> > <i>seno</i>	‘hay’
<i>ó</i> > <i>uo</i>	> <i>ů</i>	<i>kón</i> > <i>kuoň</i> > <i>kůň</i>	‘horse’
<i>’ó</i> > <i>’ie</i>	> <i>’í</i>	<i>koňóm</i> > <i>koniem</i> > <i>koním</i>	‘horse _{dat.pl} ’
<i>šč</i> > <i>št’</i>		<i>ščúr</i> > <i>štír</i>	‘scorpion’
<i>čs</i> > <i>c</i>		<i>čso</i> > <i>co</i>	‘what’

Table 1: Examples of sound/spelling changes from OC to MC

category		Old Czech	Modern Czech
infinitive		<i>péc-i</i>	<i>péc-t</i> ‘bake’
present	1sg	<i>pek-u</i>	<i>peč-u</i>
	1du	<i>peč-evě</i>	–
	1pl	<i>peč-em(e/y)</i>	<i>peč-eme</i>
imperfect	:		
	1sg	<i>peč-iech</i>	–
	1du	<i>peč-iechově</i>	–
sigm. aorist	1pl	<i>peč-iechom(e/y)</i>	–
	:		
	1sg	<i>peč-ech</i>	–
imperative	1du	<i>peč-echově</i>	–
	3du	<i>peč-esta</i>	–
	1pl	<i>peč-echom(e/y)</i>	–
imperative	:		
	2sg	<i>pec-i</i>	<i>peč</i>
	2du	<i>pec-ta</i>	–
verbal noun	2pl	<i>pec-te</i>	<i>peč-te</i>
	:		
		<i>peč-enie</i>	<i>peč-ení</i>

Table 2: A fragment of the conjugation of the verb *péci/péct* ‘bake’ (OC based on (Dostál, 1967, 74-77))

the body and the agreeing attributes. In Common Czech the dual plural distinction is completely neutralized. On the other hand, MC distinguishes animacy in masculine gender, while this distinction starts to emerge only in late OC.

2.3.3. Verbal Morphology

The system of verbal forms and constructions was far more elaborate in OC than in MC. Many forms disappeared, e.g., aorist and imperfect (simple past tenses), supine; and some became archaic, e.g., verbal adverbs, plusquamperfectum). All dual forms are no longer in MC (OC: *Herodes s Pilátem sě smřřista*; MC: *Herodes s Pilátem se smřřili* ‘Herod and Pilate reconciled’). See Table 2 for an example. The periphrastic future tense is stabilized; both *bude slůžil* and *bude slůžiti* used to mean ‘will serve’, but only the latter form is possible now.

3. Old Czech Corpus

The manuscripts and incunabula written in Old Czech are being made accessible by the Institute of Czech Language. They are transcribed and included into the Old-Czech Text Bank, which is a part of the Web Vocabulary.¹

¹See <http://vokabular.ujc.cas.cz/banka.aspx>.

So far, 124 Old Czech documents, or 2.8M tokens, have been processed and incorporated into the Old-Czech Text Bank.² Most of them date to 1400’s, the period from which most documents survived. The corpus is not balanced in respect to the periods and genres of the included documents. Nevertheless, currently, it contains a variety of documents, including liturgical, legal and medical texts, travel books, sermons, prayers, deeds, chronicles, songs, etc. Our goal is to eventually incorporate all surviving documents, including their variants. There are at least 1239 documents, as this is the number of sources of the (StčS, 1968) Old Czech dictionary.

The Old Czech spelling varied significantly. First, the period covers about 350 years, so spelling changes are expected. Second, spelling at this time was not standardized; therefore, the same word can have many different spelling variants even at the same time. Obviously, this causes many practical problems when working with the Old Czech data. For this reason, we transcribe all documents using the spelling conventions of Modern Czech, while preserving the specific features of Old Czech. This standardizes the graphemic representation of words with variant spelling, e.g., *czieřta*, *czěřta*, *cyęřta* are all represented as *čěřta*, MC: *cesta* ‘path’. It also makes the texts accessible to users without philological background. For more details, see (Lehečka and Voleková, 2010).

4. Lemmas and tagset

4.1. Principles of lemmatization

Similarly to many modern language corpora, our goal is to provide information about lemma for each word. By lemma (canonical or citation form) we mean a form distinguished from a set of all forms related by inflection. Lemmas are chosen by convention (e.g., nominative singular for nouns, infinitive for verbs). As lemmas abstract away from the inflection of words, they can be useful, for example, in searching the corpus, especially for lexicography.

However, as the language changed during the period covered in the corpus, so did lemmas. This means that the same word might be assigned different lemmas in different texts (for example, *kón*, *kuoň*, *kůň* are different historical variants of the same lemma). In some cases, a user might be interested in a particular historical variant of a lexeme, but in other they might want to search for all historical variants. As a solution, we use two levels of lemmas: (1) a traditional lemma phonologically consistent with the particular

²<http://vokabular.ujc.cas.cz/texty.aspx?id=STB>

form(s) in the text; (2) a hyperlemma, reflecting phonology around 1300. Thus, for example, the hyperlemma *kóň* would correspond to lemmas *kóň*, *kuoň*, *kůň*.

In addition, we allow a single form token to be assigned multiple lemmas and hyperlemmas and possibly, morphological tags even in a disambiguated annotation. This is used for cases when even context does not help to select a single value.

The corpus manager and viewer,³ has been modified to support these specific features of the historical corpus.

4.2. Tagset

We adopted the tag system originally developed for Modern Czech (Hajič, 2004). Every tag is represented as a string of 15 symbols each corresponding to one morphological category (2 positions out of 15 are not used). Features not applicable for a particular word have a N/A value. For example, when a word is annotated as `AAF54---2A---` it is an adjective (A), long form (A), feminine (F), singular (S), accusative (4), comparative (2), not-negated (A). The tagset has more than 4200 tags; however, only about half of them occur in a 500M token corpus.

The modification for Old Czech is quite straightforward. No additional tag positions are added, but the last slot distinguishing stylistic variants is not used. We add values for categories not present in MC (e.g., aorist, imperfect).

In addition to changes motivated by language change, we avoid using wildcard values (symbols representing a set of atomic values, e.g., H for feminine or neuter gender) for reason outlined in (Hana and Feldman, 2010). While wildcards might lead to better tagging performance, they provide less information about the word, which might be needed for linguistic analysis or an NLP application. In addition, it is trivial to translate atomic values to wildcards if needed. The Old-Czech tagset contains only wildcards covering all atomic values (denoted by X for all applicable positions). There are no wildcards covering a subset of atomic values. Forms that would be tagged with a tag containing a partial wildcard in Modern Czech are regarded as ambiguous.

5. Semi-manual lematization

We perform partial manual lemmatization of the corpus, exploiting Zipf's law (Zipf, 1935; Zipf, 1949): the 2,000 most frequent form types cover 75% of 2.8M tokens of the corpus. We manually assign lemmas to these forms, taking into account homonymy and lemma variants. The words in the corpus are then assigned candidate lemmas based on this list.

In the future, we are planning to increase the recall of this method by considering prefixes. For example *spomoci*, *přemoci*, *dopomoci* *přemoci* all have a low frequency and are thus not covered by the manually lemmatized list of frequent forms. However, they all are derived by prefixation from the word *moci* 'can', which is much more frequent and is thus covered. Also, we would like to consider regular sound change. For example, applying sound change 'ě

>e, one could translate the lemma *cesta* 'path' of *cestu* to the lemma *cěsta* of the less frequent *cěstu*.

6. Resource light morphology

The practical restrictions (no native speakers, limited corpora and lexicons, limited funding) make Old Czech an ideal candidate for the resource-light crosslingual method that we have been developing (Feldman and Hana, 2010). The first results were reported in (Hana et al., 2011). In this section, we describe the basics of our approach and some of its extensions.

The main assumption of our method (Feldman and Hana, 2010) is that a model for the target language can be approximated by language models from one or more related source languages and that the inclusion of a limited amount of high-impact and/or low-cost manual resources is greatly beneficial. We are aware of the fact that all layers of the language have changed during the last 500+ years, including phonology and spelling, syntax and vocabulary. Even words that are still used in MC often appear with different distributions, with different declensions, with different gender, etc.

6.1. Materials

Our MC *training* corpus is a portion (700K tokens) of the Prague Dependency Treebank (PDT, Hajič et al. (2006)). The corpus contains texts from daily newspapers, business and popular scientific magazines. It is manually morphologically annotated.

Several steps (e.g., lexicon acquisition) of our method require a plain text corpus. We used texts from the Old-Czech Text Bank. The corpus is significantly smaller than the corpora we used in other experiments (e.g., 39M tokens for Czech or 63M tokens for Catalan (Feldman and Hana, 2010)).

A small portion (about 1000 words) of the corpus was manually annotated for testing purposes.

6.2. Tools

6.2.1. Tagger

We use TnT (Brants, 2000), a second order Markov Model tagger. The language model of such a tagger consists of emission probabilities (corresponding to a lexicon with usage frequency information) and transition probabilities (roughly corresponding to syntax rules with strong emphasis on local word-order). We approximate the emission and transition probabilities by those trained on a modified corpus of a related language.

6.2.2. Resource-light Morphological Analysis

The *Even* tagger described in the following section relies on a morphological analyzer. While it can use any analyzer, to stay within a resource light paradigm, we use our resource-light analyzer (Hana, 2008; Feldman and Hana, 2010), which relies on a small amount of manually or semi-automatically encoded morphological details. In addition to modules we used for other languages, we also include an analyzer for Modern Czech which is used as a safety-net in parallel to an ending-based guesser.

³See <http://sourceforge.net/projects/corpmn/> for the current version

The results of the analyzer are summarized in Table 3. They show a similar pattern to the results we have obtained for other fusional languages. As can be seen, morphological analysis without any filters (the first two columns) gives good recall but also very high average ambiguity. When the automatically acquired lexicon and the longest-ending filter (analyses involving the longest endings are preferred) are used, the ambiguity is reduced significantly but recall drops as well. As with other languages, even for OC, it turns out that the drop in recall is worth the ambiguity reduction when the results are used by our MA-based taggers.

Lexicon & leo	no		yes	
	Recall	Ambiguity	Recall	Ambiguity
Overall	96.9	14.8	91.5	5.7
Nouns	99.9	26.1	83.9	10.1
Adjectives	96.8	26.5	96.8	8.8
Verbs	97.8	22.1	95.6	6.2

Table 3: Evaluation of the morphological analyzer on Old Czech

6.3. Experiments

We describe three different taggers:

1. a TnT tagger using modified MC corpus as a source of both transition and emission probabilities (section 6.3.1.);
2. a TnT tagger using modern transitions but approximating emissions by a uniformly distributed output of a morphological analyzer (MA) (sections 6.2.2. and 6.4.); and
3. a combination of both (section 6.5.).

6.3.1. Translation Model

Modernizing OC and Aging MC We modify the MC corpus so that it looks more like the OC just in the aspects relevant for morphological tagging. These modifications include translating the tagset, reversing phonological/graphemic changes, etc. Unfortunately, even this is not always possible or practical. For example, historical linguists usually describe phonological changes from old to new, not from new to old.⁴ In addition, it is not possible to deterministically translate the modern tagset to the older one. So, we modify the MC training corpus to look more like the OC corpus (the process we call ‘aging’) and also the target OC corpus to look more like the MC corpus (‘modernizing’).

Creating the Translation Tagger Below we describe the process of creating a tagger. As an example we discuss the details for the *Translation* tagger. Figure 1 summarizes the discussion.

1. Aging the MC training (annotated) corpus:

⁴Note that one cannot simply reverse the rules, as in general, the function is not a bijection.

- MC to OC tag translation:
Dropping animacy distinction (OC did not distinguish animacy).
- Simple MC to OC form transformations:
E.g., modern infinitives end in *-t*, OC infinitives ended in *-ti*;
(we implemented 3 transformations)

2. Training an MC tagger. The tagger is trained on the result of the previous step.
3. Modernizing an OC plain corpus. In this step we modernize OC forms by applying sound/graphemic changes such as those in Table 1. Obviously, these transformations are not without problems. First, the OC-to-MC translations do not always result in correct MC forms; even worse, they do not always provide forms that ever existed. Sometimes these transformations lead to forms that do exist in MC, but are unrelated to the source form. Nevertheless, we think that these cases are true exceptions from the rule and that in the majority of cases, these OC translated forms will result in existing MC words and have a similar distribution.
4. Tagging. The modernized corpus is tagged with the aged tagger.
5. Reverting modernizations. Modernized words are replaced with their original forms. This gives us a tagged OC corpus, which can be used for training.
6. Training an OC tagger. The tagger is trained on the result of the previous step. The result of this training is an OC tagger.

		Transl	Even	TranslEven
All	Full:	70.6	67.7	74.1
	SubPOS	88.9	87.0	90.6
Nouns	Full	63.1	44.3	57.0
	SubPOS	99.3	88.6	91.3
Adjs	Full:	60.3	50.8	60.3
	SubPos	93.7	87.3	93.7
Verbs	Full	47.8	74.4	80.0
	SubPOS	62.2	78.9	86.7

Table 4: Performance of various tagging models on major POS categories (in % on full tags and the SubPOS position).

The results of the translation model are provided in Table 4 (across various POS categories). The Translation tagger is already quite good at predicting the POS, SubPOS (Detailed POS) and number categories. The most challenging POS category is the category of verbs and the most difficult feature is case. Based on our previous experience with other fusional languages, getting the case feature right is always challenging. Even though case participates in syntactic agreement in both OC and MC, this category is more idiosyncratic than, say, person or tense. Therefore, the MC

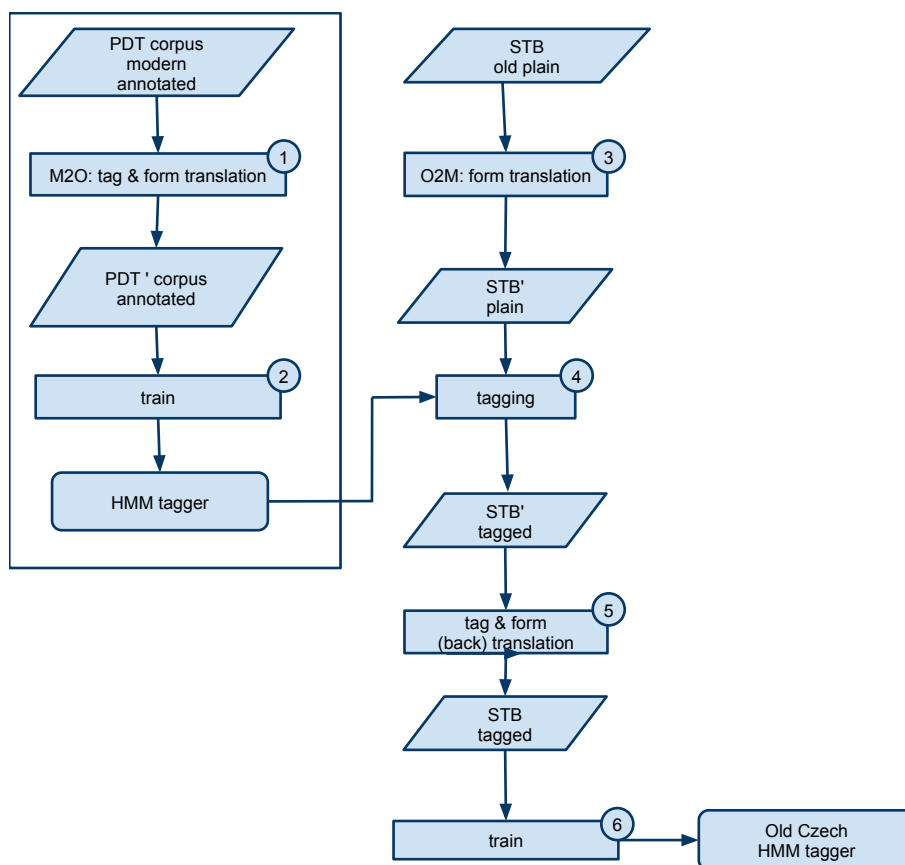


Figure 1: Schema of the Translation Tagger

syntactic and lexical information provided by the translation model might not be sufficient to compute case correctly. One of the solutions that we explore in this paper is approximating the OC lexical distribution by the resource-light morphological analyzer (see section 6.4.).

While most nominal forms and their morphological categories (apart from dual) survived in MC, OC and MC departed in verbs significantly. Thus, for example, three OC tenses disappeared in MC and other tenses replaced them. These include the OC two aorists, supinum and imperfectum. The transgressive forms are almost not used in MC anymore either. Instead MC has periphrastic past, periphrastic conditional and also future. In addition, these OC verbal forms that disappeared in MC are unique and non-ambiguous, which makes it even more difficult to guess if the model is trained on the MC data. The tagger, in fact, has no way of providing the right answer. In the subsequent sections we use the morphological analyzer described above to address this problem. Recall that our morphological analyzer uses only very basic hand-encoded facts about the target language.

6.4. Even Tagger

The *Even* tagger (see Figure 2) approximates emissions by uniformly (evenly) distributing the tags output by our morphological analyzer. The transition probabilities are based on the Aged Modern Czech corpus (result of step 2 of Figure 1). This means that the transitions are produced during the training phase and are independent of the tagged text.

However, the emissions are produced by the morphological analyzer on the basis of the tagged text during tagging.

The overall performance of the *Even* tagger drops down, but it improves on verbs significantly. Intuitively, this seems natural, because there is relatively small homonymy among many OC verbal endings (see Table 2 for an example) so they are predicted by the morphological analyzer with low or even no ambiguity.

6.5. Combining the Translation and Even Taggers

The *TranslEven* tagger is a combination of the Translation and Even models. The Even model clearly performs better on the verbs, while the Translation model predicts other categories much better. So, we decided to combine the two models in the following way. The Even model predicts verbs, while the Translation model predicts the other categories. The *TranslEven* Tagger gives us a better overall performance and improves the prediction on each individual position of the tag. Unfortunately, it slightly reduces the performance on nouns (see Table 4).

6.6. Discussion

OC and MC departed significantly over the 500+ years, at all language layers, including phonology, syntax and vocabulary. Words that are still used in MC are often used with different distributions and have different morphological forms from OC.

An additional difficulty of this task arises from the fact that our MC and OC corpora belong to different genres. While the OC corpus includes among others poetry, chronicles,

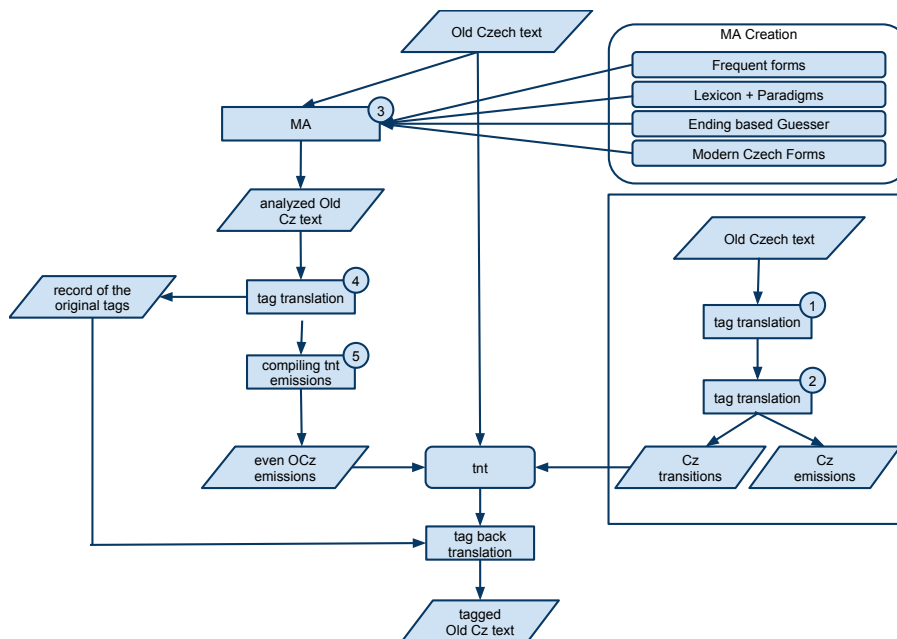


Figure 2: Schema of the MA Based Even Tagger

medical and liturgical texts, the MC corpus is mainly comprised of newspaper texts. We cannot possibly expect a significant overlap in lexicon or syntactic constructions. For example, the cookbooks contain a lot of imperatives and second person pronouns which are rare or non-existent in the newspaper texts.

Even though our tagger does not perform as the state-of-the-art tagger for Czech, the results are already useful. Remember that the tag is a combination of 12 morphological features⁵ positions out and if only one of them is incorrect, the whole positional tag is marked as incorrect. So, the performance of the tagger (74%) on the whole tag is not as low in reality. For example, if one is only interested in detailed POS (i.e., the SubPOS position, about 70 values) information the performance of our system is over 90%.

6.7. Improving morphology

To stay within the resource-light paradigm, the tagger and the morphological analyzer described in the previous section intentionally avoid resources that are unlikely to be available for a wide range of languages. However, for practical reasons it makes sense to develop tools that make use of any resources available for Old Czech.

In this section we describe a morphological analyzer improving upon the analyzer from §6.2.2. by incorporating a list of known sound changes, such as those in Table 1. We have used these changes to “translate” Old Czech words into modern Czech. Such words were analyzed by the Modern Analyzer and the result was then translated back to Old Czech. As most of the sound changes have many exceptions, we used all subsets of those rules (including an empty set), possibly assigning more than one Modern

kacřův	kacěřiev	kacřiev
kacřěv	kacěřív	kacěřév
kacieřiev	kacřív	kacěřuov
kacieřóv	kacěřův	kacieřév
kacěřóv	kacieřív	kacieřův
kacieřuov	kacřív	kacřův

Figure 3: Modernized equivalents of an Old Czech word *kacieřóv*, MC: *kacřův* ‘heretic’s’

Czech equivalent to an Old Czech word. This means the translation overgenerates, potentially assigning a number of forms exponential to the number of rules. Nevertheless, in practice this is not a problem. For example, the Old Czech word *kacřův* ‘heretic’s’ is assigned 18 different modernized equivalents (see Figure 3). However, the modern Czech analyzer recognizes only *kacřův*, the correct translation. Therefore, *kacieřóv* will be correctly analyzed as possessive adjective. Most forms have fewer than 18 modernized equivalents. The results of the analyzer incorporating a module used in such translations are given in Table 5. One can see that the results improve on nearly all categories and POS. A tagger using this analyzer achieves a similar improvement.

Lexicon & leo	no		yes	
	Recall	Ambiguity	Recall	Ambiguity
Overall	97.1	7.3	94.2	4.2
Nouns	99.0	10.0	90.6	5.8
Adjectives	96.8	17.0	96.8	8.7
Verbs	97.8	10.9	97.8	3.3

Table 5: Evaluation of the morphological analyzer using sound change rules

⁵The tag has 15 positions, but two of them are not used and we do not evaluate on the variant position as its values are to a great extent arbitrary.

6.8. Conclusion

We have presented a corpus of Old Czech currently under development. Many of the tools used during the development process are resource-light and/or rely on resources developed for Modern Czech. While the results (for example of the taggers) are significantly lower than the corresponding results for Modern Czech, they are achieved with a fraction of resources and for many practical applications are already good enough.

Acknowledgments

This research was supported by the Grant Agency of the Czech Republic (projects ID: P406/10/1140 and P406/10/P328) and by the U.S. NSF grants #0916280, #1033275. Also, this work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarín project of the Ministry of Education of the Czech Republic (project LM2010013). We would also like to thank the two anonymous reviewers for their comments.

7. References

- Thorsten Brants. 2000. TnT – A Statistical Part-of-Speech Tagger. In *Proceedings of ANLP-NAACL*, pages 224–231.
- Antonín Dostál. 1967. *Historická mluvnice česká II – Tvarosloví. 2. Časování [Historical Czech Grammar II – Morphology. 2. Conjugation]*. Praha.
- Anna Feldman and Jirka Hana. 2010. *A resource-light approach to morpho-syntactic tagging*. Rodopi, Amsterdam/New York, NY.
- Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, and Magda Ševčíková-Razímová. 2006. *Prague Dependency Treebank 2.0*. Linguistic Data Consortium, Philadelphia, PA, USA.
- Jan Hajič. 2004. *Disambiguation of Rich Inflection: Computational Morphology of Czech*. Karolinum, Charles University Press, Praha.
- Jirka Hana and Anna Feldman. 2010. A positional tagset for Russian. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 1278–1284, Valletta, Malta. European Language Resources Association.
- Jirka Hana, Anna Feldman, and Katsiaryna Aharodnik. 2011. A low-budget tagger for old czech. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 10–18, Portland, OR, USA, June. Association for Computational Linguistics.
- Jirka Hana. 2008. Knowledge- and labor-light morphological analysis. *OSUWPL*, 58:52–84.
- Laura A. Janda and Charles E. Townsend. 2002. Czech. <http://www.seelrc.org:8080/grammar/mainframe.jsp?nLanguageID=2>.
- Petr Karlík, Marek Nekula, and Zdenka Rusínová. 1996. *Příruční mluvnice češtiny [Concise Grammar of Czech]*. Nakladatelství Lidové noviny, Praha.
- Václav Křístek. 1978. *Malý staročeský slovník [Short Old Czech Dictionary]. Část Staročeské pravopisné systémy [Part Old Czech Spelling Systems]*. Praha.
- Karel Kučera. 1998. Vývoj účinnosti a složitosti českého pravopisu od konce 13. století do konce 20. století. *Slovo a slovesnost*, 59:178–199.
- Boris Lehečka and Kateřina Voleková. 2010. (Polo)automatická počítačová transkripce [(Semi)automatic computational transcription]. In *Proceedings of the Conference Dějiny českého pravopisu (do r. 1902) [History of the Czech spelling (before 1902)]*.
- Stuart E. Mann. 1977. *Czech Historical Grammar*. Hamburg: Buske.
- James Naughton. 2005. *Czech: An Essential Grammar*. Routledge, Oxon, Great Britain and New York, NY, USA.
- David Short. 1993. Czech. In Bernard Comrie and Greville G. Corbett, editors, *The Slavonic Languages*, Routledge Language Family Descriptions, pages 455–532. Routledge.
- StěS. 1968. *Staročeský slovník [Old Czech dictionary]. Část Úvodní stati, soupis pramenů a zkratek. [Part Introduction, list of sources and abbreviations]*. Praha.
- Václav Vážný. 1964. *Historická mluvnice česká II – Tvarosloví. 1. Skloňování [Historical Czech Grammar II – Morphology. 1. Declension]*. Praha.
- George K. Zipf. 1935. *The Psychobiology of Language*. Houghton-Mifflin.
- George K. Zipf. 1949. *Human Behavior and the Principle of Least-Effort*. Addison-Wesley.