A Low-budget Tagger for Old Czech

Jirka Hana
Charles University, MFF
Czech Republic
first.last@gmail.com

Anna Feldman
Montclair State University
USA
first.last@montclair.edu

Katsiaryna Aharodnik
Montclair State University
USA
ogorodnichek@gmail.com

Abstract

The paper describes a tagger for Old Czech (1200-1500 AD), a fusional language with rich morphology. The practical restrictions (no native speakers, limited corpora and lexicons, limited funding) make Old Czech an ideal candidate for a resource-light cross-lingual method that we have been developing (e.g. Hana et al., 2004; Feldman and Hana, 2010).

We use a traditional supervised tagger. However, instead of spending years of effort to create a large annotated corpus of Old Czech, we approximate it by a corpus of Modern Czech. We perform a series of simple transformations to make a modern text look more like a text in Old Czech and vice versa. We also use a resource-light morphological analyzer to provide candidate tags. The results are worse than the results of traditional taggers, but the amount of language-specific work needed is minimal.

1 Introduction

This paper describes a series of experiments in an attempt to create morphosyntactic resources for Old Czech (OC) on the basis of Modern Czech (MC) resources. The purpose of this work is two-fold. The practical goal is to create a morphologically annotated corpus of OC which will help in investigation of various morphosyntactic patterns underpinning the evolution of Czech. Our second goal is more theoretical in nature. We wanted to test the resource-light cross-lingual method that we have been developing (e.g. Hana et al., 2004; Feldman and Hana, 2010) on a source-target language pair that is divided by time instead of space. The practical restrictions (no native speakers, limited corpora and lexicons, limited funding) make OC an ideal candidate for a resource-light approach.

We understand that the task we chose is hard given the 500+ years of language evolution. We are aware of the fact that all layers of the language have changed, including phonology and graphemics, syntax and vocabulary. Even words that are still used in MC are often used with different distributions, with different declensions, with different gender, etc.

Our paper is structured as follows. We first briefly describe related work and motivate our approach. Then we outline the relevant aspects of the Czech language and compare its Modern and Old forms. Then we describe the corpora and tagsets used in our experiments. The rest of the paper describes the actual experiments, the performance of various models and concludes with a discussion of the results.

2 Related Work

Since there are no morphological taggers developed specifically for OC, we compare our work with those for MC. Morče (`http://ufal.mff.cuni.cz/morce/`) is currently the best tagger, with accuracy slightly above 95%. It is based on a statistical (averaged perceptron) algorithm which relies on a large morphological lexicon containing around 300K entries. The tool has been trained and tuned on data from the Prague Dependency Treebank (PDT; Bémova et al., 1999; Böhmová et al., 2001). The best set of features was selected after hundreds of experiments were performed. In

Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 10–18,
Portland, OR, USA, 24 June 2011. ©2011 Association for Computational Linguistics

contrast, the resource-light system we developed is not as accurate, but the amount of language-specific work needed is incomparable to that of the state-of-the-art systems. Language specific work on our OC tagger, for example, was completed in about 20 hours, instead of several years.

Research in resource-light learning of morphosyntactic properties of languages is not new. Some have assumed only partially tagged training corpora (Merialdo, 1994); some have begun with small tagged seed wordlists (Cucerzan and Yarowsky, 2002) for named-entity tagging, while others have exploited the automatic transfer of an already existing annotated resource in a different genre or a different language (e.g. cross-language projection of morphological and syntactic information as in (Cucerzan and Yarowsky, 2000; Yarowsky et al., 2001), requiring no direct supervision in the target language). The performance of our system is comparable to the results cited by these researchers.

In our work we wanted to connect to pre-existing knowledge that has been acquired and systematized by traditional linguists, e.g. morphological paradigms, sound changes, and other well-established facts about MC and OC.

3 Czech Language

Czech is a West Slavic language with significant influences from German, Latin and (in modern times) English. It is a fusional (flective) language with rich morphology and a high degree of homonymy of endings.

3.1 Old Czech

As a separate language, Czech forms between 1000-1150 AD; there are very few written documents from that time. The term Old Czech usually refers to Czech 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. Spelling during this period was not standardized, therefore the same word can have many different spelling variants. However, our corpus was transliterated – its pronunciation was recorded using the rules of the Modern Czech spelling (see Lehečka and Voleková, 2011, for more details).

change	example	
ú > *ou* non-init.	*múka* > *mouka*	‘flour’
sě > *se*	*sěno* > *seno*	‘hay’
ó > *uo* > *ů*	*kóň* > *kuoň* > *kůň*	‘horse’
šč > *šť*	*ščír* > *štír*	‘scorpion’
čs > *c*	*čso* > *co*	‘what’

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

3.2 Modern Czech

Modern Czech is spoken by roughly 10 million speakers, mostly in the Czech Republic. For a more detailed 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. Sometimes it is claimed, with some exaggeration, that it is the first foreign language the Czechs learn. The differences are mainly in phonology, morphology and lexicon. The two variants are influencing each other, resulting in a significant amount of irregularity, especially in morphology. The Czech writing system is mostly phonological.

3.3 Differences

Providing a systematic description of differences between Old and Modern Czech is well 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).

3.3.1 Phonology and Spelling

Examples of some of the more regular changes between OC and MC spelling can be found in Table 1 (Mann (1977), Boris Lehečka p.c.).

3.3.2 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. Voca-

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

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

tive is distinct only for some nouns and only in singular.

MC nouns preserved most of the features of OC, but the dual number survives only in a few paired names of parts of the body, in the declensions of the words “two” and “both” and in the word for “two hundred”. In Common Czech the dual plural distinction is completely neutralized. On the other hand, MC distinguishes animacy in masculine gender, while this distinction is only emerging in late OC.

3.3.3 Verbal Morphology

The system of verbal forms and constructions was far more elaborate in OC than in MC. Many forms disappeared all together (three simple past tenses, supinum), and some are archaic (verbal adverbs, plusquamperfectum). Obviously, all dual forms are no longer in MC. See Table 2 for an example.

4 Corpora

4.1 Modern Czech Corpus

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

The tagset (Hajič (2004)) has more than 4200 tags encoding detailed morphological information. It is a positional tagset, meaning the tags are sequences of values encoding individual morphological features and all tags have the same length, encoding all the features distinguished by the tagset. Features not applicable for a particular word have a N/A value. For example, when a word is annotated as `AAFS4----2A----` it is an adjective (`A`), long form (`A`), feminine (`F`), singular (`S`), accusative (`4`), comparative (`2`), not-negated (`A`).

4.2 Old Czech Corpora

Several steps (e.g., lexicon acquisition) of our method require a plain text corpus. We used texts from the Old-Czech Text Bank (STB, `http://vokabular.ujc.cas.cz/banka.aspx`), in total about 740K tokens. This is significantly less than we have 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. Again this is much less than what we would like to have, and we plan to increase the size in the near future. The tagset is a modification of the modern tagset using the same categories.

5 Method

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 inclusion of a limited amount of high-impact and/or low-cost manual resources is greatly beneficial and desirable.

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. Below, we describe our approach in more detail.

6 Experiments

We describe three different taggers:

1. a TnT tagger using modified MC corpus as a source of both transition and emission probabilities (section 6.1);

2. a TnT tagger using modern transitions but approximating emissions by a uniformly distributed output of a morphological analyzer (MA) (sections 6.2 and 6.3); and

3. a combination of both (section 6.4).

6.1 Translation Model

6.1.1 Modernizing OC and Aging MC

Theoretically, we can take the MC corpus, translate it to OC and then train a tagger, which would probably be a good OC tagger. However, we do not need this sophisticated, costly translation because we only deal with morphology.

A more plausible idea is to modify the MC corpus so that it looks more like the OC just in the aspects relevant for morphological tagging. In this case, the translation would include the tagset, reverse 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.[1] 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').

6.1.2 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:

    - 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.

The results of the translation model are provided in Tables 3 (for each individual tag position) and 4 (across various POS categories). What is evident from these numbers is that the Translation tagger is already quite good at predicting the POS, subPOS 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 syntactic and lexical information provided by the translation

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

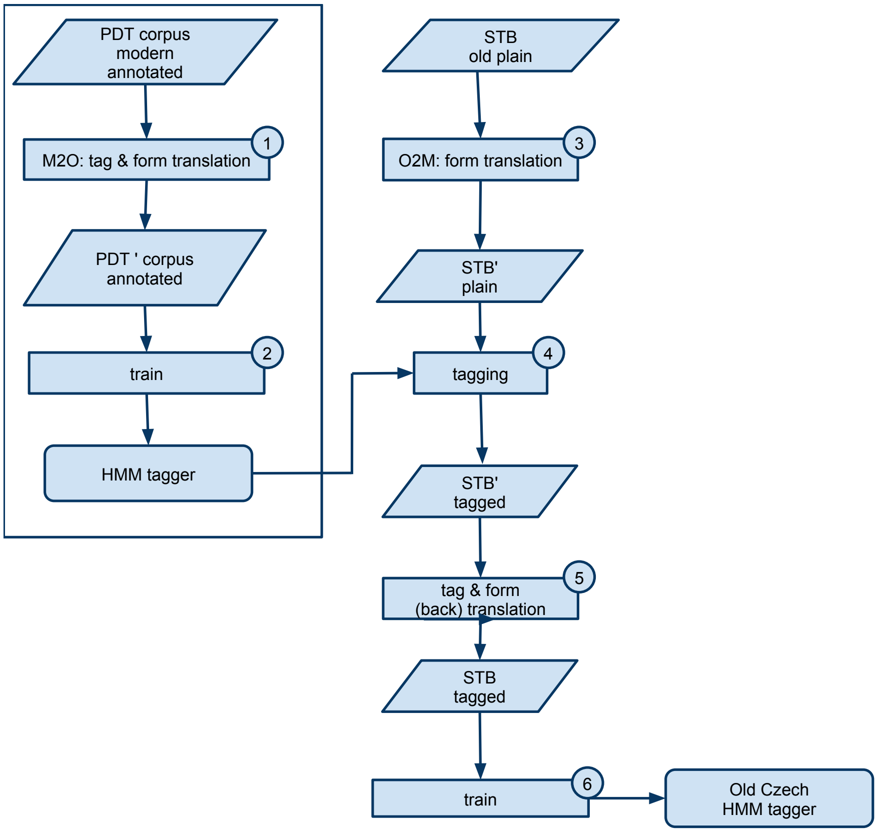

Figure 1: Schema of the Translation Tagger

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 6.3).

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 a morphological analyzer to address this problem. Our morphological analyzer uses very basic hand-encoded facts about the target language.

6.2 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 have used our resource-light analyzer (Hana, 2008; Feldman and Hana, 2010). Our approach to morphological analysis (Hana, 2008) takes the middle road between completely unsupervised systems on the one hand and systems with extensive manually-created resources on the other. It exploits Zipf's law (Zipf, 1935, 1949): not all words and morphemes matter equally. A small number of words are extremely frequent, while most words are rare. For example, in PDT, 10% most frequent noun lemmas cover about 75% of all noun tokens in the corpus. On the other hand, the less frequent 50% of noun lemmas cover only 5% of all noun tokens.

Therefore, in our approach, those resources that are easy to provide and that matter most are created

Tags:	70.6
Position 0 (POS):	91.5
Position 1 (SubPOS):	88.9
Position 2 (Gender):	87.4
Position 3 (Number):	91.0
Position 4 (case):	82.6
Position 5 (PossGen):	99.5
Position 6 (PossNr):	99.5
Position 7 (person):	93.2
Position 8 (tense):	94.4
Position 9 (grade):	98.0
Position 10 (negation):	94.4
Position 11 (voice):	95.9

Table 3: Accuracy of the Translation Model on individual positions (in %).

All	Full:	70.6
	SubPOS	88.9
Nouns	Full	63.1
	SubPOS	99.3
Adjs	Full:	60.3
	SubPos	93.7
Verbs	Full	47.8
	SubPOS	62.2

Table 4: Performance of the Translation Model on major POS categories (in %).

manually or semi-automatically and the rest is acquired automatically. For more discussion see (Feldman and Hana, 2010).

Structure The system uses a cascade of modules. The general strategy is to run “sure thing” modules (ones that make fewer errors and that overgenerate less) before “guessing” modules that are more error-prone and given to overgeneration. Simplifying somewhat the current system for OC contains the following three levels:

1. Word list – a list of 250 most frequent OC words accompanied with their possible analyses. Most of these words are closed class.

2. Lexicon-based analyzer – the lexicon has been automatically acquired from a plain corpus using the knowledge of manually provided information about paradigms (see below).

3a. Guesser – this module analyzes words relying purely on the analysis of possible endings and their relations to the known paradigms. Thus the English word *goes* would be analyzed not only as a verb, but also as plural of the potential noun *goe*, as a singular noun (with the presumed plural *goeses*), etc. In Slavic languages the situation is complicated by high incidence of homonymous endings. For example, the Modern Czech ending *a* has 14 different analyses (and that assumes one knows the morpheme boundary).

    Obviously, the guesser has low precision, and fails to use all kinds of knowledge that it potentially could use. Crucially, however, it has high recall, so it can be used as a safety net when the more precise modules fail. It is also used during lexicon acquisition, another context where its low precision turns out not to be a major problem.

3b. Modern Czech word list – a simple analyzer of Modern Czech; for some words this module gives the correct answer (e.g., *svátek* ‘holiday’, some proper names).

The total amount of language-specific work needed to provide OC data for the analyzer (information about paradigms, analyses of frequent forms) is about 12 hours and was done by a non-linguist on the basis of (Vážný, 1964; Dostál, 1967).

The results of the analyzer are summarized in Table 5. 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. Moreover, as we mentioned in the previous section, the tagger based purely on the MC corpus has no chance on verbal forms that disappeared from the language completely.

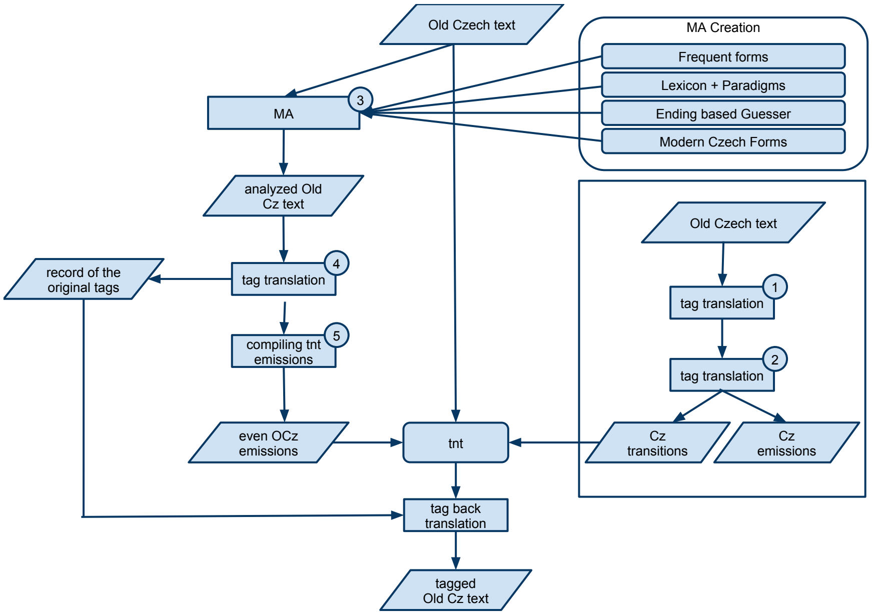

Figure 2: Schema of the MA Based Even Tagger

Lexicon & leo	no		yes	
	Recall	Ambi	Recall	Ambi
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 5: Evaluation of the morphological analyzer on Old Czech

All	Full:	67.7
	SubPOS	87.0
Nouns	Full	44.3
	SubPOS	88.6
Adjs	Full:	50.8
	SubPos	87.3
Verbs	Full	74.4
	SubPOS	78.9

Table 6: Performance of the Even Tagger on major POS categories (in %)

6.3 Even Tagger

The *Even* tagger (see Figure 2) approximates emissions by using the output of the morphological analyzer described in the previous section.

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 reason why the model is called *Even* is that the emissions are distributed evenly (uniformly; which is a crude approximation of reality).

The overall performance of the Even tagger drops down, but it improves on verbs significantly. Intuitively, this seems natural, because there is a 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.

6.4 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 Tables 7 and 8).

All	Full:	74.1
	SubPOS	90.6
Nouns	Full	57.0
	SubPOS	91.3
Adjs	Full:	60.3
	SubPos	93.7
Verbs	Full	80.0
	SubPOS	86.7

Table 7: Performance of the TranslEven tagger on major POS categories (in %)

Full tags:	74.1
Position 0 (POS):	93.0
Position 1 (SubPOS):	90.6
Position 2 (Gender):	89.6
Position 3 (Number):	92.5
Position 4 (case):	83.6
Position 5 (PossGen):	99.5
Position 6 (PossNr):	94.9
Position 7 (person):	94.9
Position 8 (tense):	95.6
Position 9 (grade):	98.6
Position 10 (negation):	96.1
Position 11 (voice):	96.4

Table 8: Performance of the TranslEven tagger on individual positions (in %).

7 Discussion

We have described a series of experiments to create a tagger for OC. Traditional statistical taggers rely on large amounts of training (annotated) data. There is no realistic prospect of annotation for OC. The practical restrictions (no native speakers, limited corpora and lexicons, limited funding) make OC an ideal candidate for a resource-light cross-lingual method that we have been developing. 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.

Additional difficulty of this task arises from the fact that our MC and OC corpora belong to different genres. While the OC corpus includes poetry, cookbooks, 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 features 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 information (the tagset that roughly corresponds to the English Penn Treebank tagset in size), the performance of our system is over 90%.

Acknowledgments

This research was generously supported by the Grant Agency Czech Republic (project ID: P406/10/P328) and by the U.S. NSF grants #0916280, #1033275, and #1048406. We would like to thank Alena M. Černá and Boris Lehečka for annotating the testing corpus and for answering questions about Old Czech. We also thank Institute of Czech Language of the Academy of Sciences of the Czech Republic for the plain text corpus of Old Czech. Finally, we thank anonymous reviewers for their insightful comments. All mistakes are ours.

References

Bémova, A., J. Hajic, B. Hladká, and J. Panevová (1999). Morphological and Syntactic Tagging of the Prague Dependency Treebank. In *Proceedings of ATALA Workshop*, pp. 21–29. Paris, France.

Böhmová, A., J. Hajic, E. Hajičová, and B. Hladká (2001). The Prague Dependency Treebank: Three-Level Annotation Scenario. In A. Abeillé (Ed.), *Treebanks: Building and Using Syntacti-*

cally Annotated Corpora. Kluwer Academic Publishers.

Brants, T. (2000). TnT – A Statistical Part-of-Speech Tagger. In *Proceedings of ANLP-NAACL*, pp. 224–231.

Cucerzan, S. and D. Yarowsky (2000). Language Independent Minimally Supervised Induction of Lexical Probabilities. In *Proceedings of the 38th Meeting of the Association for Computational Linguistics (ACL)*, Hong Kong, pp. 270–277.

Cucerzan, S. and D. Yarowsky (2002). Bootstrapping a Multilingual Part-of-speech Tagger in One Person-day. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL)*, pp. 132–138. Taipei, Taiwan.

Dostál, A. (1967). *Historická mluvnice česká II – Tvarosloví. 2. Časování [Historical Czech Grammar II - Morphology. 2. Conjugation]*. Prague: SPN.

Feldman, A. and J. Hana (2010). *A resource-light approach to morpho-syntactic tagging*. Amsterdam/New York, NY: Rodopi.

Hajič, J. (2004). *Disambiguation of Rich Inflection: Computational Morphology of Czech*. Praha: Karolinum, Charles University Press.

Hana, J. (2008). Knowledge- and labor-light morphological analysis. *OSUWPL 58*, 52–84.

Hana, J., A. Feldman, and C. Brew (2004, July). A resource-light approach to Russian morphology: Tagging Russian using Czech resources. In D. Lin and D. Wu (Eds.), *Proceedings of EMNLP 2004*, Barcelona, Spain, pp. 222–229. Association for Computational Linguistics.

Janda, L. A. and C. E. Townsend (2002). Czech.

Karlík, P., M. Nekula, and Z. Rusínová (1996). *Příruční mluvnice češtiny [Concise Grammar of Czech]*. Praha: Nakladatelství Lidové Noviny.

Lehečka, B. and K. Voleková (2011). (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)]*. in press.

Mann, S. E. (1977). *Czech Historical Grammar*. Hamburg: Buske.

Merialdo, B. (1994). Tagging English Text with a Probabilistic Model. *Computational Linguistics 20*(2), 155–171.

Naughton, J. (2005). *Czech: An Essential Grammar*. Oxon, Great Britain and New York, NY, USA: Routledge.

Short, D. (1993). Czech. In B. Comrie and G. G. Corbett (Eds.), *The Slavonic Languages*, Routledge Language Family Descriptions, pp. 455–532. Routledge.

Vážný, V. (1964). *Historická mluvnice česká II – Tvarosloví. 1. Skloňování [Historical Czech Grammar II - Morphology. 1. Declension]*. Prague: SPN.

Yarowsky, D., G. Ngai, and R. Wicentowski (2001). Inducing Multilingual Text Analysis via Robust Projection across Aligned Corpora. In *Proceedings of the First International Conference on Human Language Technology Research (HLT)*, pp. 161–168.

Zipf, G. K. (1935). *The Psychobiology of Language*. Houghton-Mifflin.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least-Effort*. Addison-Wesley.