

A Low-budget Tagger for Old Czech

Jirka Hana¹ Anna Feldman² Katsiaryna Aharodnik²

¹Charles University, Prague

²Montclair State University, NJ

ACL 2011 – LaTeCH
Portland, OR, June 24, 2010

Outline of the talk

- 1 Introduction
- 2 Czech
- 3 Corpora & Tagsets
- 4 Taggers
 - Translation Model
 - Resource-light Morphological Analysis
 - Even Tagger
 - Combining the Translation and Even Taggers
- 5 Conclusion

Outline of the talk

- 1 Introduction
- 2 Czech
- 3 Corpora & Tagsets
- 4 Taggers
 - Translation Model
 - Resource-light Morphological Analysis
 - Even Tagger
 - Combining the Translation and Even Taggers
- 5 Conclusion

Introduction

Creating morphosyntactic resources for Old Czech on the basis of Modern Czech

Two goals

- 1 **Practical:** Create morphologically annotated resources for Old Czech to investigate various morphosyntactic patterns underpinning the evolution of Czech
- 2 **Theoretical:** Test the resource-light cross-lingual method we have been developing on a source-target language pair divided by time

Difficulties

500+ years of language evolution at all layers, e.g., phonology, graphemics, syntax, vocabulary

Introduction

Creating morphosyntactic resources for Old Czech on the basis of Modern Czech

Two goals

- 1 **Practical:** Create morphologically annotated resources for Old Czech to investigate various morphosyntactic patterns underpinning the evolution of Czech
- 2 **Theoretical:** Test the resource-light cross-lingual method we have been developing on a source-target language pair divided by time

Difficulties

500+ years of language evolution at all layers, e.g., phonology, graphemics, syntax, vocabulary

Introduction

Creating morphosyntactic resources for Old Czech on the basis of Modern Czech

Two goals

- 1 **Practical:** Create morphologically annotated resources for Old Czech to investigate various morphosyntactic patterns underpinning the evolution of Czech
- 2 **Theoretical:** Test the resource-light cross-lingual method we have been developing on a source-target language pair divided by time

Difficulties

500+ years of language evolution at all layers, e.g., phonology, graphemics, syntax, vocabulary

Outline of the talk

1 Introduction

2 Czech

3 Corpora & Tagsets

4 Taggers

- Translation Model
- Resource-light Morphological Analysis
- Even Tagger
- Combining the Translation and Even Taggers

5 Conclusion

Czech

Basic info:

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

Modern Czech

- 10M speakers
- Two variants with differences mainly in phonology, morphology, lexicon
- The official variant is based on the 19th-century resurrection of the 16th century Czech
- Writing system is mostly phonological.

Old Czech

- 1150-1500 AD
- No native speakers
- Amount of available texts limited (??10MW)
- Spelling not standardized

Examples of sound/spelling changes from Old Czech to Modern Czech

change	example	
$\acute{u} > ou$ non-init.	$m\acute{u}ka > mouka$	'flour'
$s\check{e} > se$	$s\check{e}no > seno$	'hay'
$\acute{o} > uo > \acute{u}$	$k\acute{o}\check{n}\check{ } > ku\acute{o}\check{n}\check{ } > k\acute{u}\check{n}\check{ }$	'horse'
$\check{s}\check{c} > \check{s}t'$	$\check{s}\check{c}\acute{ı}r > \check{s}t\acute{ı}r$	'scorpion'
$\check{c}s > c$	$\check{c}so > co$	'what'

(Mann 1977, Boris Lehečka p.c.).

Morphology

- dual number virtually disappeared
- animacy distinction in masculine gender emerged
- many verbal forms disappeared (three simple past tenses, supinum), and some are archaic (verbal adverbs, plusquamperfectum).
- some forms have different meaning

Old vs Modern Czech verbs

category		Old Czech	Modern Czech
infinitive		peč-i	peč-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

Outline of the talk

1 Introduction

2 Czech

3 Corpora & Tagsets

4 Taggers

- Translation Model
- Resource-light Morphological Analysis
- Even Tagger
- Combining the Translation and Even Taggers

5 Conclusion

Corpora needed

Annotated corpus of Modern Czech

- PDT, 1.5M tokens.
- Daily newspapers, business and popular scientific magazines.

Plain corpus of Old Czech

- STB; <http://vokabular.ujc.cas.cz>; 740K tokens.
- Much smaller than what we used before (e.g., 63M for Catalan).
- Chronicles, legends, poetry, fiction, letters, etc.
- Transliterated.

Annotated corpus of Old Czech – for testing

- About 1000 words. Much less than we would wish for.
- Making a bigger one.

Corpora needed

Annotated corpus of Modern Czech

- PDT, 1.5M tokens.
- Daily newspapers, business and popular scientific magazines.

Plain corpus of Old Czech

- STB; <http://vokabular.ujc.cas.cz>; 740K tokens.
- Much smaller than what we used before (e.g., 63M for Catalan).
- Chronicles, legends, poetry, fiction, letters, etc.
- Transliterated.

Annotated corpus of Old Czech – for testing

- About 1000 words. Much less than we would wish for.
- Making a bigger one.

Corpora needed

Annotated corpus of Modern Czech

- PDT, 1.5M tokens.
- Daily newspapers, business and popular scientific magazines.

Plain corpus of Old Czech

- STB; <http://vokabular.ujc.cas.cz>; 740K tokens.
- Much smaller than what we used before (e.g., 63M for Catalan).
- Chronicles, legends, poetry, fiction, letters, etc.
- Transliterated.

Annotated corpus of Old Czech – for testing

- About 1000 words. Much less than we would wish for.
- Making a bigger one.

Tagset

Modern Czech

- positional tagset (Hajič 2004)
- more than 4200 tags
- encodes categories like POS, detailed POS, gender, number, case, person, voice, etc.

Old Czech

- based on the modern tagset
- roughly the same set of categories, but
- some values added (e.g. imperfect), some removed
- co-occurrence restrictions are different (e.g. dual number is not limited to few tags)

Outline of the talk

- 1 Introduction
- 2 Czech
- 3 Corpora & Tagsets
- 4 Taggers**
 - Translation Model
 - Resource-light Morphological Analysis
 - Even Tagger
 - Combining the Translation and Even Taggers
- 5 Conclusion

Modernizing OC and Aging MC

- An idea:
 - ▶ Translate an annotated MC corpus to OC; then train a tagger on the result.
 - ▶ Too costly and probably, not needed since we deal only with morphology.
- Another idea:
 - ▶ Modify the MC corpus so that it looks more like the OC just in the aspects relevant for morphological tagging.
 - ▶ Still not easy (e.g. the opposite of what historical linguistics does)
- One more idea:
 - ▶ Age the MC corpus
 - ▶ Modernize the OC corpus
 - ▶ Train on the Aged MC, tag the Modernized OC

Modernizing OC and Aging MC

- An idea:
 - ▶ Translate an annotated MC corpus to OC; then train a tagger on the result.
 - ▶ Too costly and probably, not needed since we deal only with morphology.
- Another idea:
 - ▶ Modify the MC corpus so that it looks more like the OC just in the aspects relevant for morphological tagging.
 - ▶ Still not easy (e.g. the opposite of what historical linguistics does)
- One more idea:
 - ▶ Age the MC corpus
 - ▶ Modernize the OC corpus
 - ▶ Train on the Aged MC, tag the Modernized OC

Modernizing OC and Aging MC

- An idea:
 - ▶ Translate an annotated MC corpus to OC; then train a tagger on the result.
 - ▶ Too costly and probably, not needed since we deal only with morphology.
- Another idea:
 - ▶ Modify the MC corpus so that it looks more like the OC just in the aspects relevant for morphological tagging.
 - ▶ Still not easy (e.g. the opposite of what historical linguistics does)
- One more idea:
 - ▶ Age the MC corpus
 - ▶ Modernize the OC corpus
 - ▶ Train on the Aged MC, tag the Modernized OC

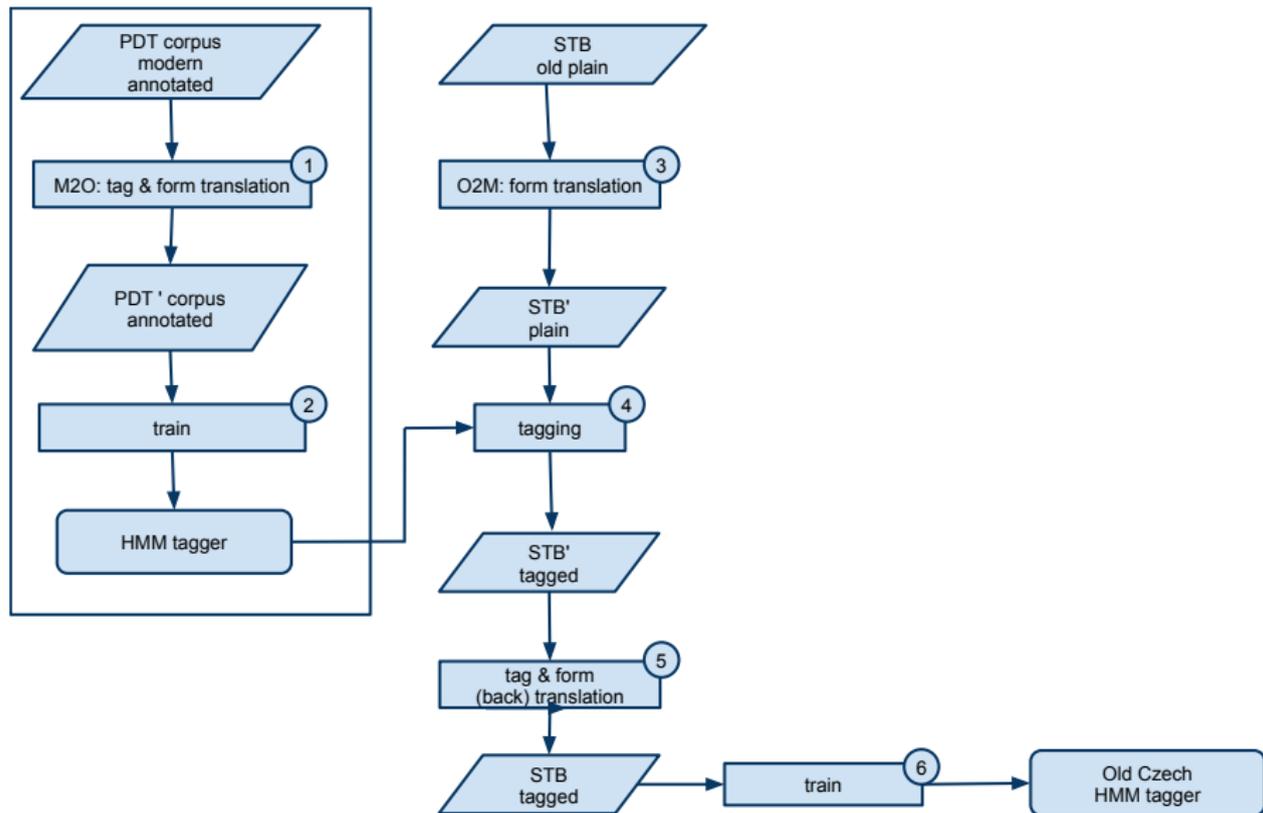
Modernizing OC and Aging MC

- An idea:
 - ▶ Translate an annotated MC corpus to OC; then train a tagger on the result.
 - ▶ Too costly and probably, not needed since we deal only with morphology.
- Another idea:
 - ▶ Modify the MC corpus so that it looks more like the OC just in the aspects relevant for morphological tagging.
 - ▶ Still not easy (e.g. the opposite of what historical linguistics does)
- One more idea:
 - ▶ Age the MC corpus
 - ▶ Modernize the OC corpus
 - ▶ Train on the Aged MC, tag the Modernized OC

Modernizing OC and Aging MC

- An idea:
 - ▶ Translate an annotated MC corpus to OC; then train a tagger on the result.
 - ▶ Too costly and probably, not needed since we deal only with morphology.
- Another idea:
 - ▶ Modify the MC corpus so that it looks more like the OC just in the aspects relevant for morphological tagging.
 - ▶ Still not easy (e.g. the opposite of what historical linguistics does)
- One more idea:
 - ▶ Age the MC corpus
 - ▶ Modernize the OC corpus
 - ▶ Train on the Aged MC, tag the Modernized OC

Translation Tagger



Translation Model – Major POSs

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

Translation Model – Individual Positions

Tags:	70.6
<hr/>	
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

Resource-light morphological analysis

- Resource-light morphological analyzer (Hana 2008, Feldman & Hana 2010)
- Manually provided information:
 - ▶ Direct analyses of frequent words
 - ▶ Endings organized into paradigms
- 12h of language-specific work needed in total. Done by a non-linguist on the basis of (Važný 1964, Dostál 1967).
- A cascade of modules:
 1. Word list – 250 most frequent words with their analyses.
 2. Lexicon-based analyzer – the lexicon has been automatically acquired from a plain corpus using the knowledge of manually provided information about paradigms.
 - 3a. Guesser – analyzes words based on their tails (string suffixes).
 - 3b. Modern Czech word list – a simple analyzer of Modern Czech;

Resource-light morphological analysis

- Resource-light morphological analyzer (Hana 2008, Feldman & Hana 2010)
- Manually provided information:
 - ▶ Direct analyses of frequent words
 - ▶ Endings organized into paradigms
- 12h of language-specific work needed in total. Done by a non-linguist on the basis of (Važný 1964, Dostál 1967).
- A cascade of modules:
 1. Word list – 250 most frequent words with their analyses.
 2. Lexicon-based analyzer – the lexicon has been automatically acquired from a plain corpus using the knowledge of manually provided information about paradigms.
 - 3a. Guesser – analyzes words based on their tails (string suffixes).
 - 3b. Modern Czech word list – a simple analyzer of Modern Czech;

Resource-light morphological analysis

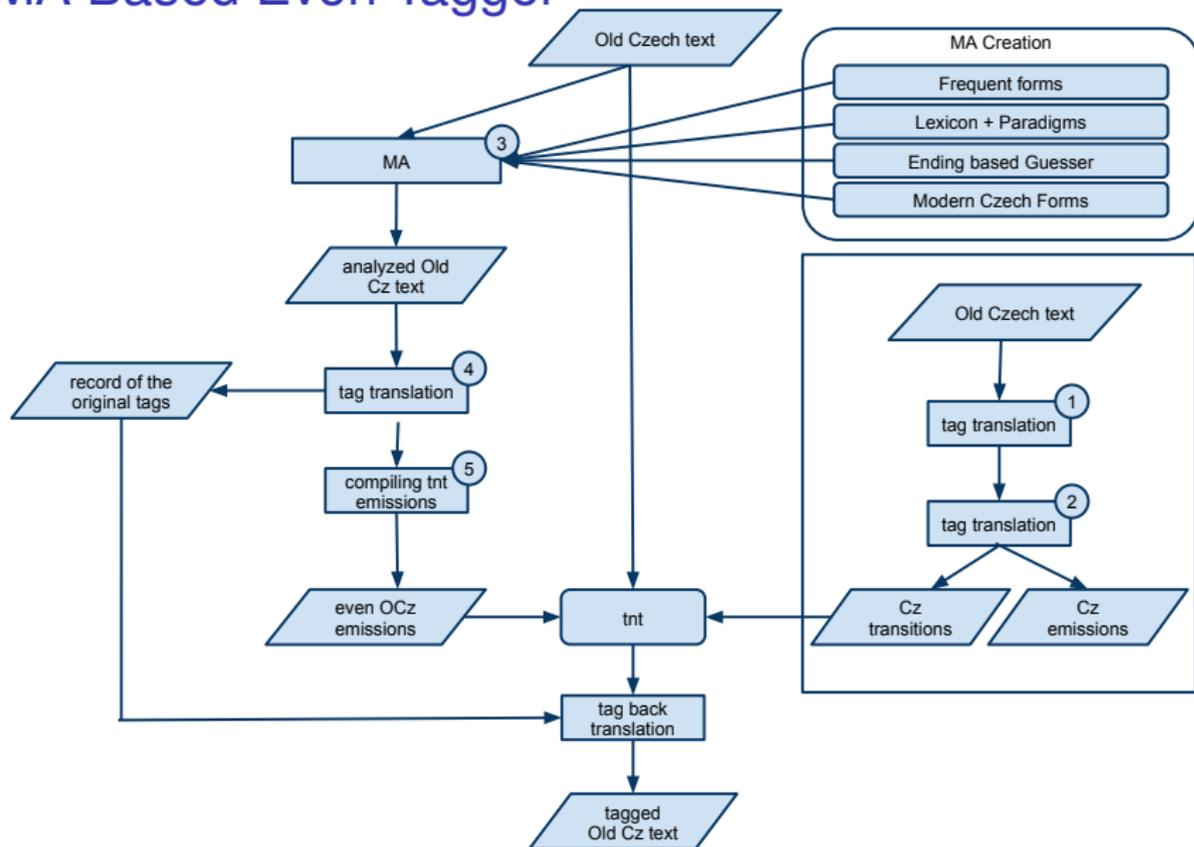
- Resource-light morphological analyzer (Hana 2008, Feldman & Hana 2010)
- Manually provided information:
 - ▶ Direct analyses of frequent words
 - ▶ Endings organized into paradigms
- 12h of language-specific work needed in total. Done by a non-linguist on the basis of (Važný 1964, Dostál 1967).
- A cascade of modules:
 1. Word list – 250 most frequent words with their analyses.
 2. Lexicon-based analyzer – the lexicon has been automatically acquired from a plain corpus using the knowledge of manually provided information about paradigms.
 - 3a. Guesser – analyzes words based on their tails (string suffixes).
 - 3b. Modern Czech word list – a simple analyzer of Modern Czech;

Resource-light morphological analysis

- Resource-light morphological analyzer (Hana 2008, Feldman & Hana 2010)
- Manually provided information:
 - ▶ Direct analyses of frequent words
 - ▶ Endings organized into paradigms
- 12h of language-specific work needed in total. Done by a non-linguist on the basis of (Važný 1964, Dostál 1967).
- A cascade of modules:
 1. Word list – 250 most frequent words with their analyses.
 2. Lexicon-based analyzer – the lexicon has been automatically acquired from a plain corpus using the knowledge of manually provided information about paradigms.
 - 3a. Guesser – analyzes words based on their tails (string suffixes).
 - 3b. Modern Czech word list – a simple analyzer of Modern Czech;

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

MA Based Even Tagger



Even Tagger on major POS categories

		Transl	
All	Full:	70.6	67.7
	SubPOS	88.9	87.0
Nouns	Full	63.1	44.3
	SubPOS	99.3	88.6
Adjs	Full:	60.3	50.8
	SubPos	93.7	87.3
Verbs	Full	47.8	74.4
	SubPOS	62.2	78.9

Ending -e and noun cases in Old Czech

case	form	lemma	gender	gloss
nom	moř-e	moře	neuter	sea
gen	oráč-e	oráč	masculine	plowman
dat	vládyc-e	vládyka	masculine	local ruler
acc	oráč-e	oráč	masculine	plowman
voc	chlap-e	chlap	masculine	guy
loc	vládyc-e	vládyka	masculine	local ruler
inst	—	—		

Old Czech verbs

category		Old Czech	Modern Czech
infinitive		peč-i	peč-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

- The Even model clearly performs better on the verbs (and pronouns, conjunctions, ...),
- The Translation model predicts other categories much better.
- Use Even for verbs etc, Translation for the rest.

Even Tagger on major POS categories

		Transl	
All	Full:	70.6	74.1
	SubPOS	88.9	90.6
Nouns	Full	63.1	57.0
	SubPOS	99.3	91.3
Adjs	Full:	60.3	60.3
	SubPos	93.7	93.7
Verbs	Full	47.8	80.0
	SubPOS	62.2	86.7

Combined tagger on individual positions

Full tags:	74.1
<hr/>	
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

Outline of the talk

- 1 Introduction
- 2 Czech
- 3 Corpora & Tagsets
- 4 Taggers
 - Translation Model
 - Resource-light Morphological Analysis
 - Even Tagger
 - Combining the Translation and Even Taggers
- 5 Conclusion**

Conclusion

- Traditional statistical taggers rely on large amounts of training data – There is no realistic prospect of annotation for Old Czech.
- Old Czech is an ideal candidate for testing our resource-light method – no native speakers, limited corpora and lexicons, limited funding
- Challenging: Old Czech and Modern Czech departed significantly over the 500+ years; Old Czech and Modern Czech corpora belong to different genres.
- Results: 74% accuracy on the whole tag, 90+% on detailed POS.

Thanks to

- the Grant Agency Czech Republic (project ID: P406/10/P328)
- the U.S. NSF grants #0916280, #1033275, and #1048406.
- Alena M. Černá and Boris Lehečka for annotating the testing corpus and for answering questions about Old Czech.
- Institute of Czech Language of the Czech Academy of Sciences for the plain text corpus of Old Czech
- Anonymous reviewers for their insightful comments