A Low-budget Tagger for Old Czech

Jirka Hana\textsuperscript{1} Anna Feldman\textsuperscript{2} Katsiaryna Aharodnik\textsuperscript{2}

\textsuperscript{1}Charles University, Prague
\textsuperscript{2}Montclair 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
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
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

<table>
<thead>
<tr>
<th>change</th>
<th>example</th>
<th>example</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>ú &gt; ou</td>
<td>múka &gt; mouka</td>
<td>‘flour’</td>
<td></td>
</tr>
<tr>
<td>sě &gt; se</td>
<td>sěno &gt; seno</td>
<td>‘hay’</td>
<td></td>
</tr>
<tr>
<td>ó &gt; uo &gt; ů</td>
<td>kóň &gt; kuoň &gt; kůň</td>
<td>‘horse’</td>
<td></td>
</tr>
<tr>
<td>šč &gt; šť</td>
<td>ščír &gt; štír</td>
<td>‘scorpion’</td>
<td></td>
</tr>
<tr>
<td>čs &gt; c</td>
<td>čso &gt; co</td>
<td>‘what’</td>
<td></td>
</tr>
</tbody>
</table>

(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

<table>
<thead>
<tr>
<th>category</th>
<th>Old Czech</th>
<th>Modern Czech</th>
<th>'bake'</th>
</tr>
</thead>
<tbody>
<tr>
<td>infinitive</td>
<td>péc-i</td>
<td>péc-t</td>
<td></td>
</tr>
<tr>
<td>present</td>
<td>1sg pek-u</td>
<td>peč-u</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1du peč-evě</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1pl peč-em(e/y)</td>
<td>peč-eme</td>
<td></td>
</tr>
<tr>
<td></td>
<td>imperfect</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1sg peč-iech</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1du peč-iechově</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1pl peč-iechom(e/y)</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>imperative</td>
<td>2sg pec-i</td>
<td>peč</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2du pec-ta</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2pl pec-te</td>
<td>peč-te</td>
<td></td>
</tr>
<tr>
<td>verbal noun</td>
<td>peč-enie</td>
<td>peč-ení</td>
<td></td>
</tr>
</tbody>
</table>
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 & Tagsets

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](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

1. PDT corpus modern annotated
   - M2O: tag & form translation
   - PDT' corpus annotated
   - train
   - HMM tagger

2. STB old plain
   - O2M: form translation
   - STB' plain
   - tagging
   - STB' tagged
   - tag & form (back) translation
   - STB tagged

3. train
4. train
5. train
6. Old Czech HMM tagger

J. Hana et al. (Charles University & MSU)  A Low-budget Tagger for Old Czech
## Translation Model – Major POSs

<table>
<thead>
<tr>
<th>POS</th>
<th>Full:</th>
<th>SubPOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>70.6</td>
<td>88.9</td>
</tr>
<tr>
<td>Nouns</td>
<td>63.1</td>
<td>99.3</td>
</tr>
<tr>
<td>Adjs</td>
<td>60.3</td>
<td>93.7</td>
</tr>
<tr>
<td>Verbs</td>
<td>47.8</td>
<td>62.2</td>
</tr>
</tbody>
</table>
Translation Model – Individual Positions

<table>
<thead>
<tr>
<th>Tags:</th>
<th>70.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position 0 (POS ):</td>
<td>91.5</td>
</tr>
<tr>
<td>Position 1 (SubPOS ):</td>
<td>88.9</td>
</tr>
<tr>
<td>Position 2 (Gender ):</td>
<td>87.4</td>
</tr>
<tr>
<td>Position 3 (Number ):</td>
<td>91.0</td>
</tr>
<tr>
<td>Position 4 (case ):</td>
<td>82.6</td>
</tr>
<tr>
<td>Position 5 (PossGen):</td>
<td>99.5</td>
</tr>
<tr>
<td>Position 6 (PossNr ):</td>
<td>99.5</td>
</tr>
<tr>
<td>Position 7 (person ):</td>
<td>93.2</td>
</tr>
<tr>
<td>Position 8 (tense ):</td>
<td>94.4</td>
</tr>
<tr>
<td>Position 9 (grade ):</td>
<td>98.0</td>
</tr>
<tr>
<td>Position 10 (negation):</td>
<td>94.4</td>
</tr>
<tr>
<td>Position 11 (voice ):</td>
<td>95.9</td>
</tr>
</tbody>
</table>
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;
<table>
<thead>
<tr>
<th>Lexicon &amp; leo</th>
<th>no</th>
<th>yes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>Ambi</td>
</tr>
<tr>
<td>Overall</td>
<td>96.9</td>
<td>14.8</td>
</tr>
<tr>
<td>Nouns</td>
<td>99.9</td>
<td>26.1</td>
</tr>
<tr>
<td>Adjectives</td>
<td>96.8</td>
<td>26.5</td>
</tr>
<tr>
<td>Verbs</td>
<td>97.8</td>
<td>22.1</td>
</tr>
</tbody>
</table>
MA Based Even Tagger

Old Czech text

MA

analyzed Old Cz text

record of the original tags

tag translation

compiling tnt emissions

even Ocz emissions

MA Creation

Frequent forms
Lexicon + Paradigms
Ending based Guesser
Modern Czech Forms

Old Czech text

tag translation

tag translation

tnt

cz transitions

cz emissions

tag back translation

tagged Old Cz text
### Even Tagger on major POS categories

<table>
<thead>
<tr>
<th>Category</th>
<th>Full:</th>
<th>SubPOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>70.6</td>
<td>67.7</td>
</tr>
<tr>
<td></td>
<td>88.9</td>
<td>87.0</td>
</tr>
<tr>
<td>Nouns</td>
<td>63.1</td>
<td>44.3</td>
</tr>
<tr>
<td></td>
<td>99.3</td>
<td>88.6</td>
</tr>
<tr>
<td>Adjs</td>
<td>60.3</td>
<td>50.8</td>
</tr>
<tr>
<td></td>
<td>93.7</td>
<td>87.3</td>
</tr>
<tr>
<td>Verbs</td>
<td>47.8</td>
<td>74.4</td>
</tr>
<tr>
<td></td>
<td>62.2</td>
<td>78.9</td>
</tr>
</tbody>
</table>
## Ending -e and noun cases in Old Czech

<table>
<thead>
<tr>
<th>case</th>
<th>form</th>
<th>lemma</th>
<th>gender</th>
<th>gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>nom</td>
<td>moř-e</td>
<td>moře</td>
<td>neuter</td>
<td>sea</td>
</tr>
<tr>
<td>gen</td>
<td>oráč-e</td>
<td>oráč</td>
<td>masculine</td>
<td>plowman</td>
</tr>
<tr>
<td>dat</td>
<td>vládyc-e</td>
<td>vládyka</td>
<td>masculine</td>
<td>local ruler</td>
</tr>
<tr>
<td>acc</td>
<td>oráč-e</td>
<td>oráč</td>
<td>masculine</td>
<td>plowman</td>
</tr>
<tr>
<td>voc</td>
<td>chlap-e</td>
<td>chlap</td>
<td>masculine</td>
<td>guy</td>
</tr>
<tr>
<td>loc</td>
<td>vládyc-e</td>
<td>vládyka</td>
<td>masculine</td>
<td>local ruler</td>
</tr>
<tr>
<td>inst</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>
## Old Czech verbs

<table>
<thead>
<tr>
<th>category</th>
<th>Old Czech</th>
<th>Modern Czech</th>
</tr>
</thead>
<tbody>
<tr>
<td>infinitive</td>
<td>péč-i</td>
<td>péč-t</td>
</tr>
<tr>
<td>present</td>
<td></td>
<td>‘bake’</td>
</tr>
<tr>
<td>1sg</td>
<td>pek-u</td>
<td>peč-u</td>
</tr>
<tr>
<td>1du</td>
<td>peč-evě</td>
<td>–</td>
</tr>
<tr>
<td>1pl</td>
<td>peč-em(e/y)</td>
<td>peč-eme</td>
</tr>
<tr>
<td>imperfect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1sg</td>
<td>peč-iech</td>
<td>–</td>
</tr>
<tr>
<td>1du</td>
<td>peč-iechově</td>
<td>–</td>
</tr>
<tr>
<td>1pl</td>
<td>peč-iechom(e/y)</td>
<td>–</td>
</tr>
<tr>
<td>imperative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2sg</td>
<td>pec-i</td>
<td>peč</td>
</tr>
<tr>
<td>2du</td>
<td>pec-ta</td>
<td>–</td>
</tr>
<tr>
<td>2pl</td>
<td>pec-te</td>
<td>peč-te</td>
</tr>
<tr>
<td>verbal noun</td>
<td>peč-enie</td>
<td>peč-ení</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Category</th>
<th>Transl Full</th>
<th>Transl SubPOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>70.6</td>
<td>74.1</td>
</tr>
<tr>
<td></td>
<td>88.9</td>
<td>90.6</td>
</tr>
<tr>
<td>Nouns</td>
<td>63.1</td>
<td>57.0</td>
</tr>
<tr>
<td></td>
<td>99.3</td>
<td>91.3</td>
</tr>
<tr>
<td>Adjs</td>
<td>60.3</td>
<td>60.3</td>
</tr>
<tr>
<td></td>
<td>93.7</td>
<td>93.7</td>
</tr>
<tr>
<td>Verbs</td>
<td>47.8</td>
<td>80.0</td>
</tr>
<tr>
<td></td>
<td>62.2</td>
<td>86.7</td>
</tr>
</tbody>
</table>
## Combined tagger on individual positions

<table>
<thead>
<tr>
<th>Position</th>
<th>Tag Type</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>POS</td>
<td>93.0</td>
</tr>
<tr>
<td>1</td>
<td>SubPOS</td>
<td>90.6</td>
</tr>
<tr>
<td>2</td>
<td>Gender</td>
<td>89.6</td>
</tr>
<tr>
<td>3</td>
<td>Number</td>
<td>92.5</td>
</tr>
<tr>
<td>4</td>
<td>case</td>
<td>83.6</td>
</tr>
<tr>
<td>5</td>
<td>PossGen</td>
<td>99.5</td>
</tr>
<tr>
<td>6</td>
<td>PossNr</td>
<td>94.9</td>
</tr>
<tr>
<td>7</td>
<td>person</td>
<td>94.9</td>
</tr>
<tr>
<td>8</td>
<td>tense</td>
<td>95.6</td>
</tr>
<tr>
<td>9</td>
<td>grade</td>
<td>98.6</td>
</tr>
<tr>
<td>10</td>
<td>negation</td>
<td>96.1</td>
</tr>
<tr>
<td>11</td>
<td>voice</td>
<td>96.4</td>
</tr>
<tr>
<td><strong>Full tags</strong></td>
<td></td>
<td><strong>74.1</strong></td>
</tr>
</tbody>
</table>
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