Strategies for automatic morphological tagging of non-standard Czech

David Lukeš, Zuzana Komrsková
{david.lukes,zuzana.komrskova}@ff.cuni.cz
Roadmap

Objectives

1. to characterize the specificities of informal spoken Czech transcripts contained in the ORAL series corpora, as compared with standard written Czech

2. based on this, to devise ways of improving the performance of morphological taggers on this data

See also lukes+15

Introduction

Specificities of spoken data

Methodology

Challenges
Introduction
Informal spoken language

- language as used in the **family circle**, among **friends** and close people in general
- **spontaneous** (× prepared in advance)
- **minimal self-consciousness** about the formal attributes of one’s speech
- inevitable consequences for the **quality of the recordings**
Transcription and processing

- transcription guidelines consciously reflect orality:
  - morphological and lexical variation
  - no sentence boundaries in ORAL2013.
- speech transcripts vs. written-text-based NLP tools—two approaches:
  - focus on information extraction (using a pre-existing NLP pipeline)? → adapt (normalize) transcript
  - focus on linguistic description of spoken language? → adapt tools
Corpora of informal spoken Czech (available via https://korpus.cz)

<table>
<thead>
<tr>
<th>corpus</th>
<th>tokens</th>
<th>positions</th>
<th>time span</th>
<th>region</th>
<th>hrs of audio</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMK</td>
<td>674,992</td>
<td>819,267</td>
<td>1988–1996</td>
<td>Prague</td>
<td>N/A</td>
</tr>
<tr>
<td>BMK</td>
<td>500,460</td>
<td>596,009</td>
<td>1994–1999</td>
<td>Brno</td>
<td>N/A</td>
</tr>
<tr>
<td>ORAL2006</td>
<td>1,000,798</td>
<td>1,312,282</td>
<td>2002–2006</td>
<td>west of the country</td>
<td>111</td>
</tr>
<tr>
<td>ORAL2008</td>
<td>1,000,097</td>
<td>1,349,536</td>
<td>2002–2007</td>
<td>west of the country</td>
<td>115</td>
</tr>
</tbody>
</table>

**Sociolinguistic metadata**: sex, age group, highest achieved level of education, (region of origin)
Regional distribution of speakers in the ORAL2006 corpus
Regional distribution of speakers in the ORAL2008 corpus
Regional distribution of speakers in the ORAL2013 corpus
Specificities of spoken data
Token-level differences from written text

- additional **homonymy**, **out-of-vocabulary** word forms
  - **spoken** language variants
    - **protože** (*because*) \(\rightarrow\) *poče, potože, pže, prče, proe, …* (OOV)
    - **jsem** (*to be, 1\textsuperscript{st} pers. sg. pres.*) almost universally pronounced and transcribed as *sem*, homonymous with adv. *sem* (*here*)
  - **regional** variants
    - n. *kámen* (*stone*) \(\rightarrow\) regional *kameň*, homonymous with IMP. of v. *kamenět* (*to turn to stone*)

- solutions
  - manually **extend dictionary** to account for OOV forms ✓
  - **vowel length** and **palatalization** alternations \(\sim\) **diacritics** ✓
Structural differences from written text

Excerpt of multi-party interaction from the ORAL2013 corpus, one speaker per line.

- non-trivial context retrieval ⇒ broken syntactic dependencies
  - turn unit split to account for overlap (can be fixed) ⇒ orphaned object (governed by head, ← above)
  - completion of syntactic structure by other speaker (much harder to detect) ⇒ orphaned modifier (agreement with head, ← above)
Structural differences from written text II

XML corpus pseudo-source corresponding to excerpt in previous slide.
Methodology
Iterative improvement workflow leveraging the speed of the **MorphoDiTa** tagging framework (strakova+14). Original morphological dictionary and training data: **MorfFlex CZ, PDT 3.0** (hajic+90; straka+13).
## Hand-curated list of changes

### "zaděj" a "vzaděj" jako komparativ od <vzadu>
> vzadu Dg-------2A---- vzaděj
> vzadu Dg-------2A---- zaděj

### "ni" ani "ňu" nemá být spojka, jen zájmeno
< ni_,a J^--------------

### "ňu" může být kromě zájmene ještě interjekce
> ňu II-------------- ňu

### "pote" jen jako imperativ od <jít>, ne jako vokativ od <pot>
< pot NNIS5------A---- pote
> jít Vi-P---2--A---1 pote
II Simple non-standard → standard mapping

abysem    abych
abysme    abychom
abyzme    abychom
ádný      žádné
akorat    akorát
akorát    akorát
akurát    akorát
### Rule-generated variants

```perl
## rozgenerovat "tej", "jednej" apod. z té, jedné apod.
gen $entry, tag => '..FS.*', word => '.*é',
    from => 'é$', to => 'ej';
## "nedělu", "chvilu", "kruháču" apod.
gen $entry, tag => '..FS4.*', word => '.*[\$vow]i',
    from => 'i$', to => 'u';
```
IV Phonetic variants

bavit_:T VB-P---1P-AA--- bavime
bavit_:T Vt-P---3P-AA--2 baviť
bavit_:T Vt-S---3P-AA--2 baviť
...
žvýkačka_^(*2) NNFP2------A---- žvykaček
žvýkačka_^(*2) NNFP1------A---- žvykačky
žvýkačka_^(*2) NNFP4------A---- žvykačky
žvýkačka_^(*2) NNFP5------A---- žvykačky
žvýkačka_^(*2) NNFS2------A---- žvykačky
Challenges
Challenges

▶ what is the “right” lemma/tag anyway?
  ▶ **univerbation**: 
    - (pro)sim tě vs. (pro)simtě
  ▶ level of **lemma abstraction**:
    - separate lemmas for forms with **v-prothesis**?
    - \{tek’a, tekkom, tek’kon, tek’ko, tě\} ⊆ lemma **Ted’** or not?
    - similarly with the prolific variation in **reinforced demonstratives**: tuten, tadyten, henten, tenhleten, tendleten, tenhlecten …
  ▶ **semantic bleaching**: vole (\textit{voc. of noun vůl} → phatic/expressive particle)

▶ many subtly different **project-specific transcription norms**
▶ no **gold standard**
Hand-annotating a gold standard
Selected bibliography 1
Thank you for your attention!
This presentation was supported by the Czech National Corpus project (LM2011023) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

https://korpus.cz