Experimental Tagging of the ORAL Series Corpora: Insights on Using a Stochastic Tagger

David Lukeš, Petra Klimešová, Zuzana Komrsková and Marie Kopřivová
Institute of the Czech National Corpus, Faculty of Arts, Charles University in Prague

CZECH NATIONAL CORPUS

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

Introduction

- 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

ORAL	size tokens	positions	time span	regional coverage	hours
2006	1,000,798	1,312,282	2002–2006	west of the country	111
2008	1,000,097	1,349,536	2002–2007	west of the country	115
2013	2,785,189	3,285,508	2008–2011	entire country	292
total	**4,786,084**	**5,947,326**	**2002–2011**	**entire country**	**518**

Table 1: The ORAL series corpora of informal spoken Czech: **private conversations between family and friends**. Transcription guidelines consciously reflect orality: **morphological** and **lexical** variation, **no sentence boundaries** in ORAL0213.

Method

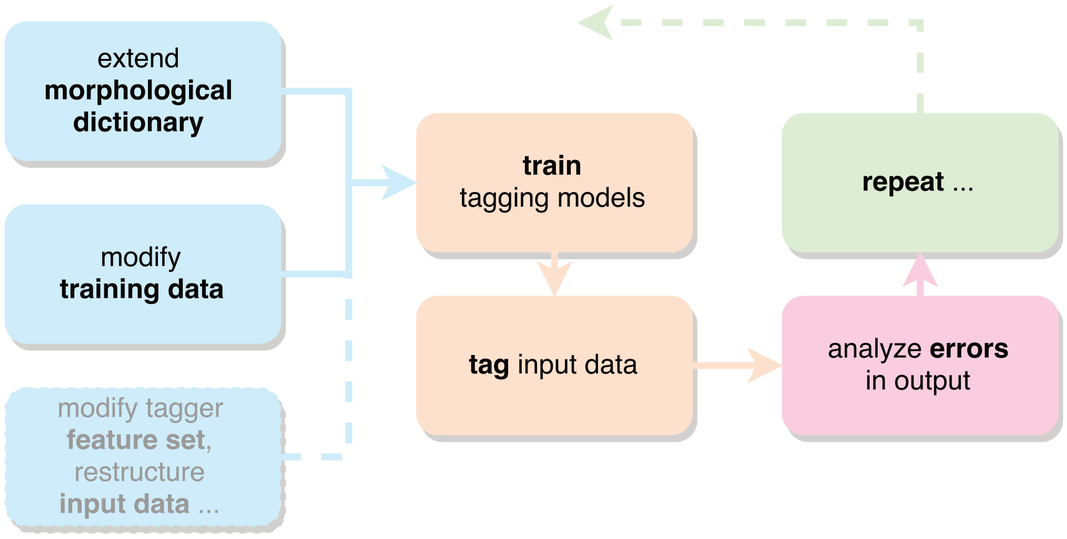

Figure 1: Iterative improvement workflow leveraging the speed of the **MorphoDiTa** tagging framework. Original morphological dictionary and training data: **MorfFlex CZ**, **PDT 3.0**.

Token-level differences from written text

- additional **homonymy**, **out-of-vocabulary** word forms
  - **spoken** language variants
    - *protože* (*because*) → *poče*, *potože*, *pže*, *prče*, *proe*, ... (OOV)
    - *jsem* (*to be*, 1ST PERS. SG. PRES.) almost universally pronounced and transcribed as *sem*, homonymous with adv. *sem* (*here*)
  - **regional** variants
    - n. *kámen* (*stone*) → 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 ∼ **diacritics** ⇒ **remove non-standard** ones and use existing software to **automatically add standard** ones as a pre-processing step ✗

Structural differences from written text

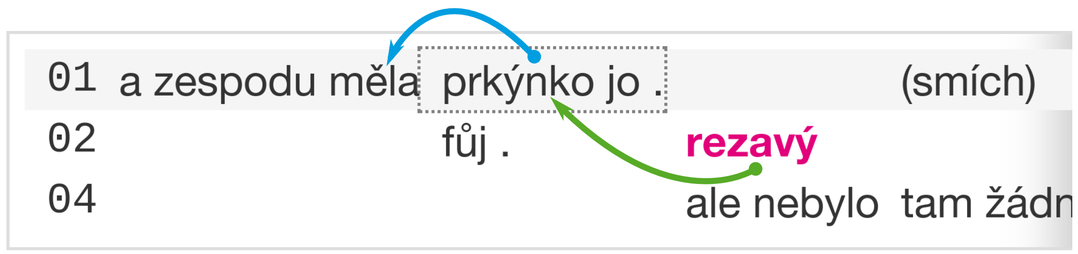

Figure 2: Excerpt of multi-party interaction from the ORAL2013 corpus, one speaker per line.

- non-trivial **context retrieval** ⇒ broken **syntactic dependencies**
  - turn unit split to account for **overlap** (can be fixed) ⇒ orphaned **object** (**governed** by head, ← in Fig. 2)
  - **completion** of syntactic structure **by other speaker** (much harder to detect) ⇒ orphaned **modifier** (**agreement** with head, ← in Fig. 2)

```
<sp num="01">a zespodu měla</sp>
<sp num="02" overlap="true">fůj .</sp>
<sp num="01" overlap="true">prkýnko jo .</sp>
<sp num="02" overlap="true">rezavý</sp>
```

Figure 3: XML corpus pseudo-source corresponding to excerpt in Fig. 2.

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**?
    - {*teďka*, *teďkom*, *teďkon*, *teďko*, *teď*} ⊂ lemma TEĎ or not?
    - similarly with the prolific variation in **reinforced demonstratives**: *tuten*, *tadyten*, *henten*, *tenhleten*, *tendleten*, *tenhlecten* ...
  - **semantic bleaching**: *vole* (VOC. of noun *vůl* → phatic/expressive particle)
- many subtly different **project-specific transcription norms**
- no gold standard

Hand-annotating a gold standard (in progress)

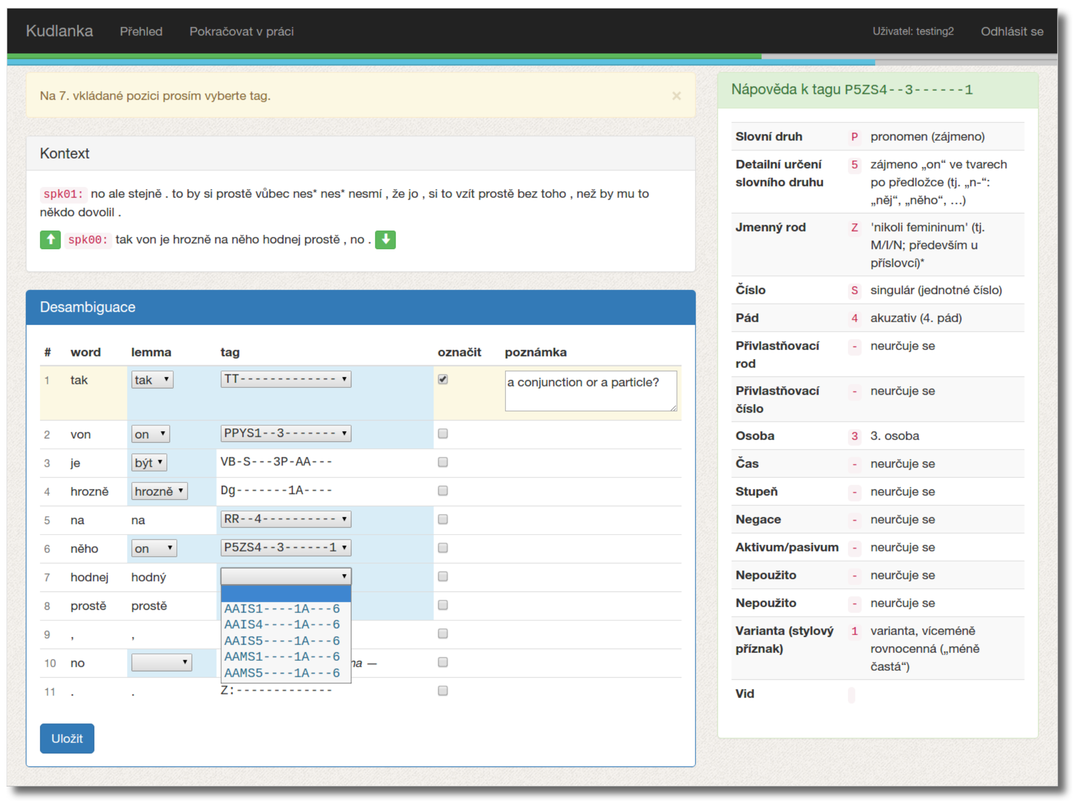

Figure 4: Kudlanka, an on-line manual disambiguation interface. The UI includes an adaptive disambiguation form (blue box), expandable context (gray box), tag hints (green box) and asynchronous error feedback (yellow box). See `https://github.com/dlukes/kudlanka`.

Acknowledgments

This presentation is backed 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**.

CZECH NATIONAL CORPUS
http://korpus.cz
{lukes,klimesova,komrskova,kopriva}@korpus.cz