ÚFAL Research

Ondřej Bojar, Vincent Kríž
{bojar,kriz}@ufal.mff.cuni.cz
Ústav formální a aplikované lingvistiky
Matematicko-fyzikální fakulta
Univerzita Karlova v Praze

Outline of Our Presentation

- Linguistic Tools and Data.
- NLP Applications.
- Previews and Demos.

Linguistic Tools

- Language and character encoding identification.
- Finding boundaries of sentences and words:

  Švejk 12. prosince dorazil na král. Vinohrady s dopisem.

  /'aIs.kri:m/ → I scream / icecream.
- Morphological analysis.
- Surface and deep syntactic analysis.
- Named entity recognition.

  The White House said...

  Rice University ≠ university of rice
- Coreference (includes finding what pronouns refer to)

Linguistic Data

- **Corpora** are (large) collections of text:
  - Often with annotation and/or sentence structure:
    Prague Dependency Treebank (PDT): 1.5M words.
    Prague Czech-English Dep. Tbk. (PCEDT): 50k sentence.
  - Some multilingual:
    CzEng (15M sentences, 220M words, ∼50 metres of books)
    HindEnCorp (274k sentences, 3.8M words)
- **Dictionaries** by ÚFAL are machine readable:
  - Morphological dict.: *kočka* is a Czech word and *kočke* is not.
  - Valency dict. says that:

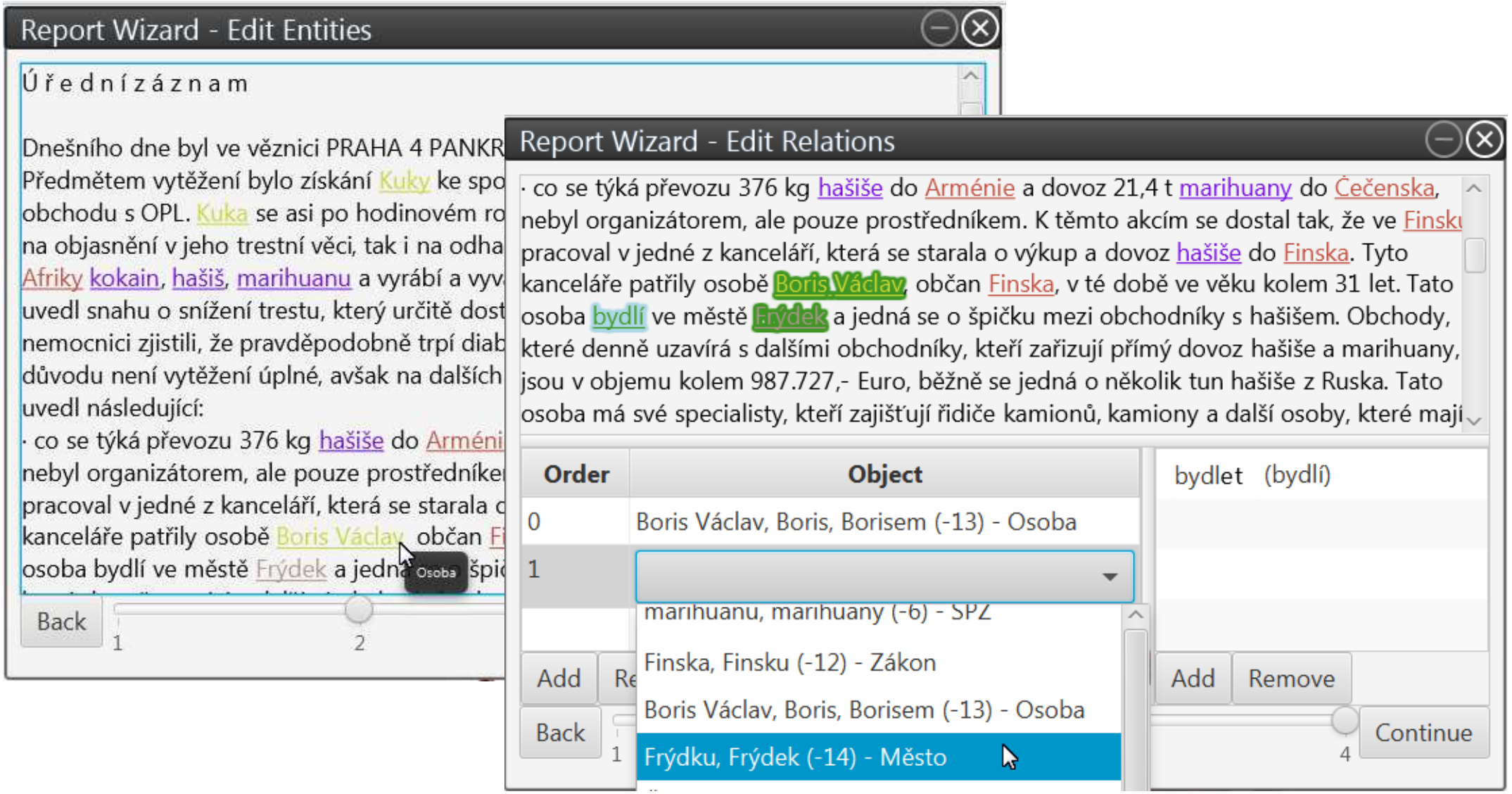
    Rodiče přijali Petra. → is a correct Czech sentence

    Rodiče přijeli Petra. → is not
  - Subjectivity lexicon lists evaluative expressions (good, bad, ...).

Linguistic Applications

	UFAL does this Well	UFAL does this Currently
Spell Checker	**	*
Grammar Checker	**	
Information (Document) Retrieval	partners	Khresmoi
Information Extraction	*, partners	Khresmoi
Automatic Summarization	partners	Khresmoi
Speech Recognition and Synthesis	*	**
Dialogue Systems	*	Vystadial
Machine Translation	**	MosesCore, Faust
Speech Translation	*	AMALACH
Sentiment Analysis (Opinion Mining)	*	**

Content Analysis



Mock screenshots from student software project Textan.

Random Ideas for Cooperation

Implementation is ready, just launch:

- Making text data „denser“:
  - Conversion of word forms to their basic forms.
  - Finding content words, removal of auxiliary words.
    ⇒ Makes keyword spotting or topic detection easier.
- Sentence analysis: subject, predicate, object. Pronoun resolution.
- Detection of named entities (*U tří slunců*), collocations (*bílý kůň*).
- Document classification and clustering (politics, sport, e-mail, ...).

Random Ideas for Future Tools:

- Social network monitoring.
- Authorship identification.

Links

ÚFAL:

`http://ufal.mff.cuni.cz/`

→ Research → Prague Czech-English Dependency Treebank 2.0

→ Data: Sample Czech and English surface and deep syntactic trees

→ Video Recordings

→ Online tools and demos: `http://ufal.mff.cuni.cz/tools/`

LINDAT:

`http://ufal.mff.cuni.cz/lindat/`

- Repository of linguistic data.
- We can run „annotations on demand“.