[DISC] Discovery issues

Eleri Aedmaa, Chloé Artaud, Goranka Blagus Bartolec,
Guillaume Chiron, Vittorio Ganfi, Daniela Majchráková,
Justina Mandravickaite, Shiva Taslimipoor
Supervisor: Fabienne Cap

the 2nd PARSEME Training School in La Rochelle, 2016

Issues

- Token- vs. type-based identification of MWEs
- Conflating variants of the same MWE.
- How to apply MWEs for post-OCR correction?

Token- vs. type-based identification of MWEs

- Expressions which are MWEs in some contexts
- Shorter MWEs are more problematic
- Example: *have children*
  - She had her child in this hospital. (idiomatic: giving birth)
  - A: Do you have children? B: Yes, I have two.
- Croatian: *dignuti ruke*
  - word for word translation: to raise hands(acc. pl.)
  - idiomatic meaning: 'surrender, give up'
  - literal meaning: to raise hands (e.g. during training)

Token- vs. type-based identification of MWEs

- Solutions:
  - Some previous work on word sense disambiguation
  - Detecting MWEs in context
    - Extracting semantic and syntactic features from context
    - How large should the context be?
  - Classifying sentences that include expressions

Conflating variants of the same MWE.

- **Problem**: How automatically detect that various abbreviations or acronyms are the same MWE?
- **Example**: French for "overtime" in salary slips:
  - *Heures supplémentaires*
  - *H. Supp.*
  - *Heures suppl*
  - *Heures supplém.*
  - *HS*
- **Solutions**:
  - Use of regular expressions
  - Extraction of complete MWE from Google Result Page

MWEs for post-OCR correction?

- **Problem**:
  - Poor recognition rate (<70%)
  - Automatic approach needed
  - Small annotated corpora
- **Hypothesis**: OCRs tend to reproduce same mistakes
- **Paradigm**: Machine translation problem (decoding). Learn to talk the "OCR output" language.
- **Solution**:
  - N-gram model,
  - MWEs model, even misspelled ones (no semantics attached)
  - Levenshtein distance

Questions?

- Token- vs. type-based identification of MWEs
- Conflating variants of the same MWE.
- How to apply MWEs for post-OCR correction?