EmpiriST 2.0: Adding normalization, lemmatization and semantic tags to a German web and social media corpus

Thomas Proisl · Natalie Dykes · Philipp Heinrich · Besim Kabashi · Stefan Evert
Lehrstuhl für Korpus- und Computerlinguistik, FAU Erlangen-Nürnberg

Data Set

- Ongoing annotation with UCREL Semantic Analysis System (USAS)
- 232 tags in 21 discourse fields:
  - F: Food and farming
  - I: Money and commerce
    - F1: Food
      - 0-11 (Kinder Egg)
      - Geschmackserlebnis (taste adventure)
    - F2: Drinks
      - Kaffee (coffee), Bierdose (beer can)

Semantic tagging

- Preliminary inter-annotator agreement: 56.9 for fine-grained tags and 73.8 for discourse fields; fortunately disagreements are largely systematic
  - Prepositions: location/direction or grammatical marker?
  - Emotions and action words "freu" ("happy"): emotion/action + speech act?
  - User names (marc30, Nudelsuppensterne): Z1 Personal names or Z3 Other names?

Example sentence

```
<posting id="cmc_train_003_099" author="quaki" origid="#-114">
</s>
```

Lemmatization and off-the-shelf tools

- Lemmatization guidelines based on TIGER, extensions for new POS tags in STTS_IBK
  - Contractions treated like APPRART: Use lemma of first constituent
  - Many new tags cover tokens that do not inflect anyway
- Two lemmatization strategies: Surface-oriented lemmatization and normalized lemmatization
  - Surface-oriented lemmata
    - Mainly based on inflectional suffixes
    - Retain, as far as possible, non-standard orthographical features
      - Griffe → Griff, hinstellt → hinstellen
    - Inter-annotator agreement: 96.04–96.86
  - Normalized lemmata
    - Based on normalized word forms
    - Create, as far as possible, standard language lemmata
      - Griffe → Griff, hinstellt → hinstellen
    - Inter-annotator agreement: 95.88–96.67

Lemmatization baselines and off-the-shelf tools

- Accuracy 93.85–94.45 due to guideline changes w.r.t. proper names
- Creative spellings:
  - Elisions:
    - müllsäcke: müllsäcke
  - Typos
    - forskrafted: forskrafted
    - pleite: pleite
- Creative spellings:
  - Elisions:
    - hund: hund
  - Typos
    - geil: geil

Normalization

- CMC data often deviate from norms of written language → Conceptually closer to spoken language
- Affects syntax, lexical choices, spelling
  - Contractions: gehts (= geht es), some (= so eine), …
  - Elisions: ne (= eine), hinziehn (= hinziehen), …
- Creative spellings: verStichl (= verdreißachte), …
- Emphasis via character repetitions: daaaaaa (= da), geeee (= geil)
- Typos
- Our normalization efforts:
  - Correct obvious typos: das/dass, hinstellt → hinstellt, Griffe → Griff
  - Normalize to post-spelling-reform orthography: mus → muss
  - Normalize non-lexicalized forms to established standard forms: hund → Hund, zB → z.B., usw → u.h, nen → einen, Disku → Diskussion
- Independent normalization by four student helpers
- Inter-annotator agreements: 98.02–98.23 (Cohen’s $\kappa$)
- Accuracy 93.85–94.45 due to guideline changes w.r.t. proper names

http://www.linguistik.fau.de
https://github.com/fau-klue/empirist-corpus
<firstname>.-<lastname>@fau.de