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

- **EmpiriST 2015 gold standard** (Beißwenger et al. 2016)
  - Shared task on automatic annotation of computer-mediated communication (CMC) and web corpora
    - **CMC**: tweets, social and professional chats, comments, wiki talk pages
    - **Web**: web sites, blogs, Wikipedia articles, Wikinews
  - Manually tokenized & annotated – STTS_IBK
  - STTS + 18 additional tags (Beißwenger et al. 2015)
  - Sentence boundaries and UD POS tags by Rehbein et al. (2018)
  - Converted to vertical format suitable for CQPweb, SketchEngine, etc.

Normalization

- CMC data often deviate from norms of written language → Conceptually closer to spoken language
- Affects syntax, lexical choices, spelling
  - Contractions: **geht** (= geht es), sone (= so eine), …
  - Elisions: ne (= eine), hinziehn (= hinziehen), …
  - Creative spellings: verStache (= verdreifachte), …
  - Emphasis via character repetitions: daahaaa (= da), geeeeil (= geil)
- Typos
- Our normalization efforts:
  - Correct obvious typos: das/dass, hinstell → hinstellt, Griffe → Griff
  - Normalize to post-spelling-reform orthography: muß → muss
  - Normalize non-lexicalized forms to established standard forms: hund → Hund, z.B. → z.B., usuh → uh, 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

Lemmatization

- 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, hinstell → 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, hinstell → hinstellen
  - Inter-annotator agreement: 95.88–96.67

Semantic tagging

- Ongoing annotation with UCREL Semantic Analysis System (USAS)
- 232 tags in 21 discourse fields:
  - **F**: Food and farming
    - F1: Food
    - F2: Drinks
  - **I**: Money and commerce
    - I1: Money generally
    - I2: Money: affluence
  - 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, Nudelsuppenster): Z1 Personal names or Z3 Other names?

Example sentence

```xml
<posting id="en_train_001_001" author="quaki" origid="1-114">
<s>
  die ART DET DET 25 die der der
  vischea NN NOUN 22 Vischer Visch Viech
  reissen VVFIN VERB A1.1.2 reißen reissen reißen
  imma ADV ADV T1.1 immer immer
  die ART DET DET 25 die der der
  müllsäcke NN NOUN 02 Müllsäcke Müllsack Müllsack
  auf PTINF PART A1D auf auf auf
  bitten 11C 115T 24 bis bittet bitten
</s>
</posting>
```

- 7 columns: Word form, STTS IBK tag, UD POS tag, USAS tag, normalized form, surface-oriented lemma, normalized lemma

Lemmatization baselines and off-the-shelf tools

![Performance of baseline systems and off-the-shelf tools](image.png)

- **Task**: Ongoing annotation with UCREL Semantic Analysis System (USAS)
- **Strategy**: Ongoing annotation with UCREL Semantic Analysis System (USAS)
- **Performance**: 96.04–96.86