FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

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

	CMC	Web
Training	5,109	4,944
Test	5,237	7,568
Total	10,346	12,512

- 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: *gehts* (= *geht es*), *sone* (= *so eine*), . . .
  - Elisions: *ne* (= *eine*), *hinziehn* (= *hinziehen*), . . .
  - Creative spellings: *ver3fachte* (= *verdreifachte*), . . .
  - Emphasis via character repetitions: *dahaaaa* (= *da*), *geeeil* (= *geil*)
  - Typos
- Our normalization efforts:
  - Correct obvious typos: *das/dass*, *hinstelt* → *hinstellt*, *Grigfe* → *Griffe*
  - Normalize to post-spelling-reform orthography: *muß* → *muss*
  - Normalize non-lexicalized forms to established standard forms: *hund* → *Hund*, *zB* → *z.B.*, *uuuh* → *uh*, *nen* → *einen*, *Disku* → *Diskussion*
- Independent normalization by four student helpers
- Inter-annotator agreements: 98.02–98.23 (Cohen's κ)
- 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
  - *Grigfe* → *Grigf*, *hinstelt* → *hinstelen*
  - Inter-annotator agreement: 96.04–96.86
- Normalized lemmata
  - Based on normalized word forms
  - Create, as far as possible, standard language lemmata
  - *Grigfe* → *Griff*, *hinstelt* → *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*
  `Ü-Ei` (Kinder Egg), `Geschmackserlebnis` (taste adventure)
- F2: *Drinks*
  `Kaffee` (coffee), `Bierdose` (beer can)

I: **Money and commerce**
- I1: *Money generally*
  `Euro`, `zahlen` (pay)
  - I1.1: *Money: affluence*
    `arm` (poor)
  - I1.2: *Money: debts*
    `pleite` (broke)

- 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?
  - Emoticons and action words **freu** (*happy*): emotion/action + speech act?
  - User names (*marc30*, *Nudelsuppenstern*): Z1 *Personal names* or Z3 *Other names*?

Example sentence

```
<posting id="cmc_train_003_099" author="quaki" origid="1-114">
<s>
die         ART    DET  Z5     die         der       der
viecha      NN     NOUN L2     Viecher     Viech     Viech
reissen     VVFIN  VERB A1.1.2 reißen      reissen   reißen
imma        ADV    ADV  T1.1   immer       imma      immer
die         ART    DET  Z5     die         der       der
müllsäcke   NN     NOUN O2     Müllsäcke   Müllsack  Müllsack
auf         PTKVZ  PART A10    auf         auf       auf
hmmmm       ITJ    INTJ Z4     hm          hmmmm     hm
</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