**A German Twitter Snapshot**

Tatjana Scheffler  
Universität Potsdam  
tatjana.scheffler@uni-potsdam.de  
@tschfflr

### Motivation
- more than 800,000 German tweets/day
- but only < 1% of all tweets are German
- previous analyses mostly on English data

### Gardenhose corpora
- pick percentage of all tweets at random -> biased sampling?
- tweets out of context -> discourse features cannot be observed

**Goal:** (Almost) Complete snapshot of German-language Twitter data

### German Twitter Corpus
- collected April 1-30, 2013
- tweets: 24,179,189
- distinct users: 1,907,891
- geo-tagged tweets: 263,364 (< 1.1%)
- replies: 5,133,544 (21.2%)
- top ten clients (spam removal): 19,258,112 (79.6%)

### Corpus construction
- Twitter stream
- keyword tracking
- language filter

- German stop word list
- langID

- German Twitter corpus

### Corpus completeness
- keyword list recall
  - parallel tracking of keyword list, geolocation-based stream and a frequent user list
  - four days: 1.8 mio tweets through keyword, 365K location, 30k user (follow) stream
  - evaluate how many of the location and follow streams were also recalled by keyword tracking:
    - location: 97.2%
    - follow (user list): 94.6%

- language filtering
  - manual evaluation on a small subset of preliminary data
  - error analysis shows complementary errors of langID and Google CLD modules

**German Tweets | langID | Google CLD | Twitter**
| Precision | 97% | 96% | ~ 40% |

### Twitter Threads
- over 30% of tweets are part of a conversation
- in_reply_to_id creates discussion trees

### private conversations
- few participants
- length = depth of thread

### celebrity broadcasts
- many participants
- depth of thread ≈ 2

Depth vs. length of all threads on April 1, 2013:

### German Twitter Users
- unique users: 1,907,891
- u. users in geo-tagged tweets: 46,559
- most-tweeting “users”: over 28,500 tweets

### spam removal
- users in threads more likely to be real:
  - avg. tweets/user: 12.7
  - avg. tweets/user (replies): 5.7
- restrict clients:
  - top-ten clients: 79.6% of tweets
  - small clients often bots’ APIs

### Deliberation: Twitter Causes
- causal connectives are frequent in Twitter:
  - 1.7% of tweets / 2.6% of replies
  - “spoken”/informal style of justification

Relative frequencies of connectives ‘denn’, ‘da’, and ‘nämlich’ compared with ‘weil’ (all, ‘because’) in corpora of spoken and written German, and in Twitter.

**Twitter** = Wulff-corpus; 253,172 German tweets about the Wulff-scandal // **bmp** = Berliner Morgenpost // **FOLK** = Forschungs- und Lehrkorpus Gesprochenes Deutsch // **Wegener** = spoken corpora 1980-1999 from (Wegener 1999, Tab. 1) // **Rudolph** = written texts (Rudolph 1982) referenced in (Wegener 1999)

For Twitter and FOLK, the frequencies of causal ‘denn’ and ‘da’ were estimated by manually disambiguating a representative sample of the data. 0 values = no data

### Acknowledgement

Project: Analyse von Diskursen in Sozialen Medien, funded by BMBF, # 01UG1232A
Web: http://www.social-media-analytics.org/

9th Language Resources and Evaluation Conference (LREC), May 26–31, 2014, Reykjavik, Iceland