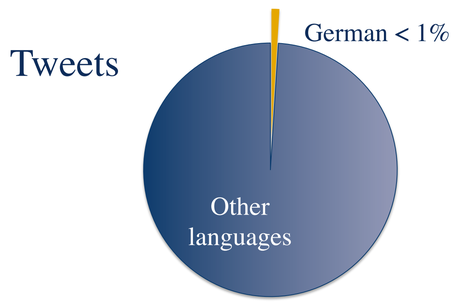
A German Twitter Snapshot

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

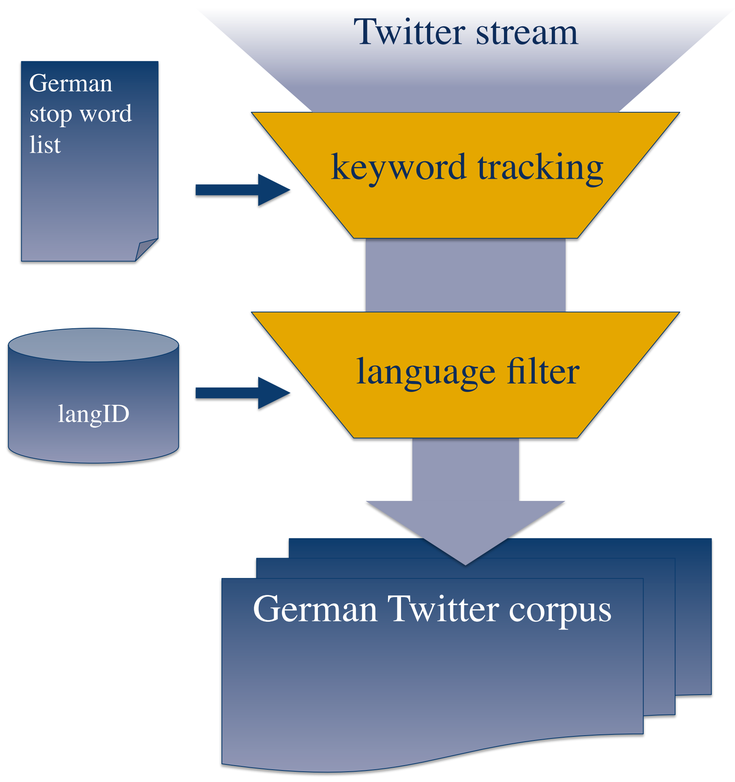
Goal: (Almost) Complete snapshot of German-language Twitter data

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



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%

rate limiting

- rate limiting becoming more of a problem for foreign language tweets
- up to 4.5 mio tweets missed by rate limit
- only small fraction of those (16%) are actually German
- < 3% loss due to rate limits

Total Missed Data: ~10%

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

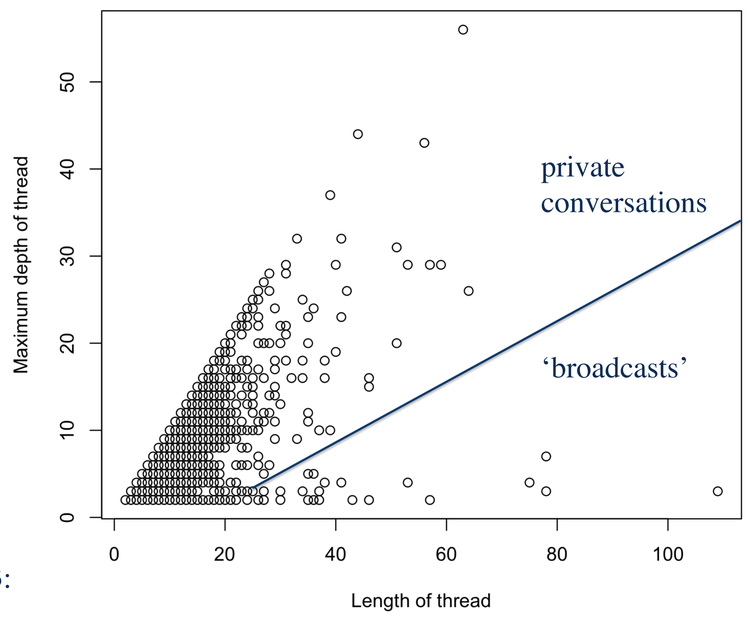
- over 30% of tweets are part of a conversation
- `in_reply_to_id` creates discussion trees

private conversations

- few participants
- length ≈ depth of thread

celebrity broadcasts

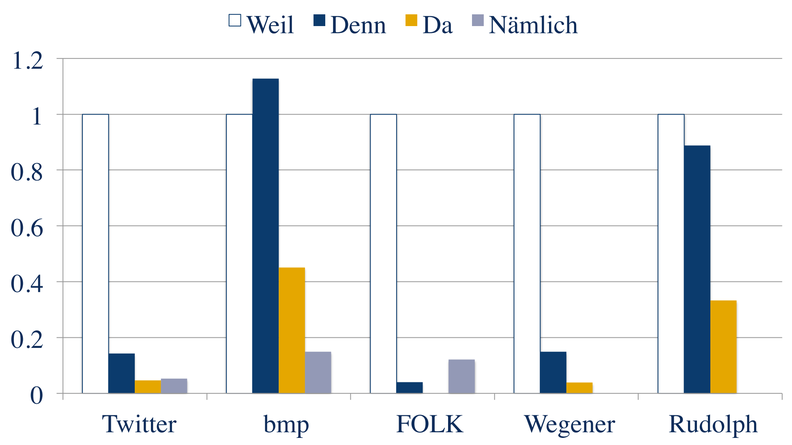
- many participants
- depth of thread ≈ 2



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

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/COSMAS II (daily newspaper) // **FOLK** = Forschungs- und Lehrkorpus Gesprochenes Deutsch; dialogs // **Wegener** = spoken corpora 1980-1999 from (Wegener 1999, Tab. 1) // **Rudolph** = written texts (Rudolph 1982) referenced in (Wegener 1999)

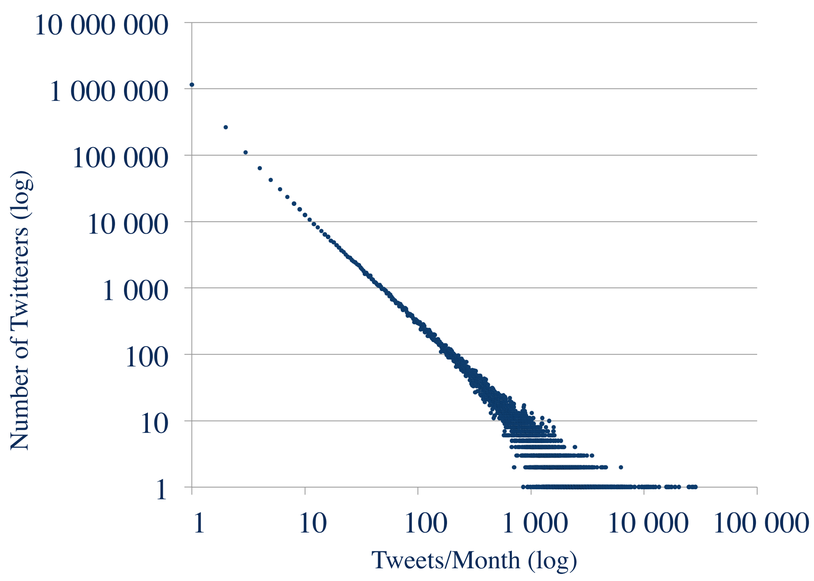
For Twitter and FOLK, the frequencies of causal 'denn' and 'da' were estimated by manually disambiguating a representative sample of the data. 0 values = no data

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



Acknowledgement

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