Deutsche Twitterdaten – Korpuserstellung und Analyse

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

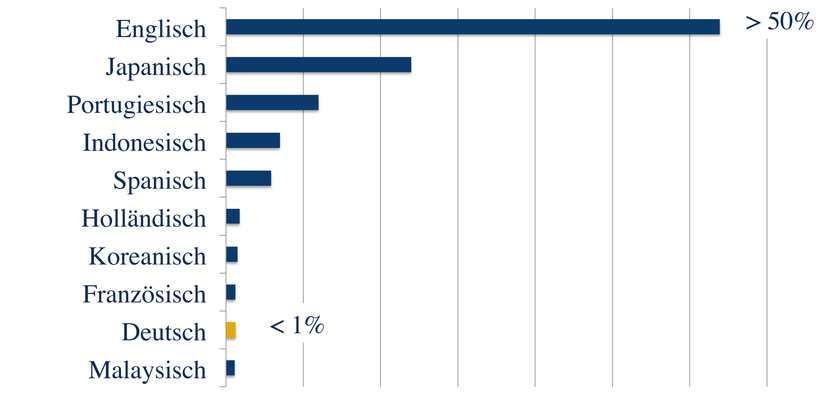
Motivation

- Mindestens 400.000 deutsche Tweets pro Tag
  = ca. 6 Mio. Wörter pro Tag
- z.Vgl.: IDS Referenzkorpora: > 5 Mrd. Wörter

- Bisherige Analysen fast ausschließlich auf englischen Daten
- In der computerlinguistischen Praxis wichtig: Trenderkennung und –analyse, Tonalitätsanalysen, Profiling, Social Media Monitoring
- Auch als Quelle für linguistische Analysen interessant: große Datenmenge, leicht erhältlich, informeller Sprachstil, Registererffekte (s. Beispiel unten Mitte)

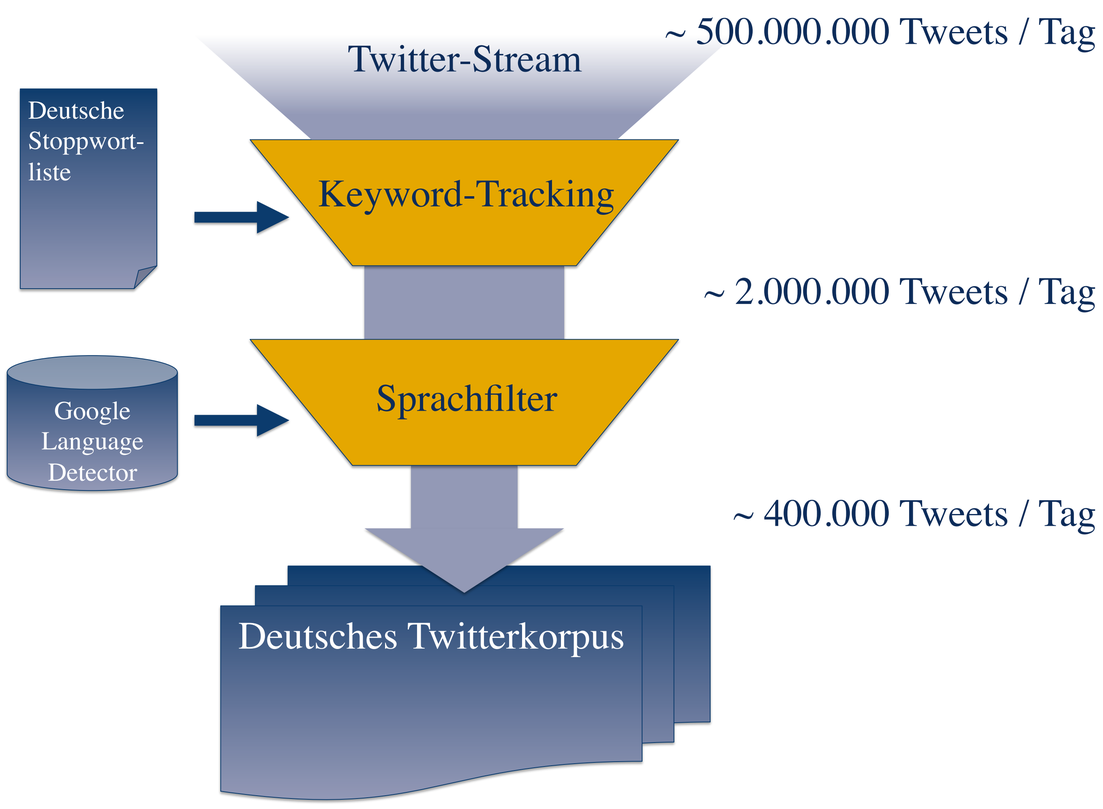
- **Ziel**: Möglichst komplette Sammlung deutscher Tweets über einen Zeitraum

Sprache auf Twitter



Quelle: Hong, Lichan, Convertino, Gregorio, and Chi, Ed. "Language Matters In Twitter: A Large Scale Study" International AAAI Conference on Weblogs and Social Media (2011)

Korpuserstellung



Bisher verfügbare Daten:

2.-12. Dezember 2011	ca. 4,5 Mio. Tweets
20.-27. Dezember 2011	ca. 4 Mio. Tweets
Dezember/Januar 2013	> 8 Mio. Tweets

Anleitung und Tools

Twitter-Stream mitschneiden

1. Python-Paket: tweepy `https://github.com/tweepy/tweepy`
2. Eigene Anwendung bei Twitter registrieren und Access/Consumer Keys erhalten
3. Wortliste der mitzuschneidenden Stichwörter erstellen
   - Z.B.: Filtere Stream nach 397 häufigen deutschen Wörtern
   - Ausschluss von fremdsprachigen Homographen: “war”, “die”, “des”, …
   - Verlust nur ca. 2-5% der deutschen Tweets
4. Twitter für Linguisten-Paket Twython starten
   `http://www.ling.uni-potsdam.de/~scheffler/twitter/`

Sprachidentifikation

- Twitter-eigene Sprachklassifikation ist zu inakkurat; scheint auf Eigenschaften im User-Profil zu basieren
- Google Compact Language Detector `http://pypi.python.org/pypi/chromium_compact_language_detector/`
- Langid `https://github.com/saffsd/langid.py`
  nach Forschung von Liu und Baldwin “langid.py: An Off-the-shelf Language Identification Tool” (ACL 2012)

Deutsche Tweets	Langid	Google CLD	Twitter
Präzision	97%	96%	~ 40%

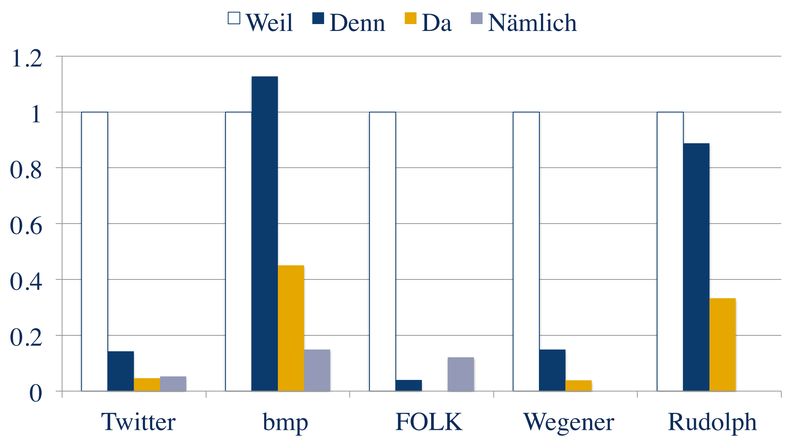
Twitterdaten

- Potenziell nützliche Metadaten
  - Nutzerinformationen (Ort, Name, Anzahl Follower und Freunde)
  - Antwortrelation, Retweets
  - Ortsinformationen (nur bei < 2% der deutschen Tweets)
- Spezielle Tokens (Emoticons, URLs, # Hashtags: Themenmarkierungen)
- Umgangssprache, Slang und Dialekte
- **Vorverarbeitung**
  - Normalisierung (Umlaute, Prolongationen, Tippfehler?)
  - Behandlung von Spezialtokens
  - Tokenisierung
  - Satzgrenzenbestimmung

☐ uuund der akku hält und hält.... :) #iphone4s

☐ Der Tagesspiegel: Busemann: Keine Weisung an Staatsanwaelte in Wulffff-Affaere - http://t.co/Xef3vrUj #Pressemitteilung

Beispiel “Twitter-Stil”: Kausalkonnektoren



Auftretenshäufigkeiten von “denn”, “da” und “nämlich” relativ zu “weil” in gesprochenen und geschriebenen Korpora, sowie auf Twitter.
Twitter = Wulff-Korpus; 253172 Deutsche Tweets über den Wulff-Skandal
bmp = Berliner Morgenpost-Teil von COSMAS II
FOLK = Forschungs- und Lehrkorpus Gesprochenes Deutsch; Dialoge
Wegener = Gesprochene Korpora 1980-1999 aus (Wegener 1999, Tab. 1)
Rudolph = Geschriebene Texte (Rudolph 1982) zitiert in (Wegener 1999)
Bei Twitter und FOLK wurden die Häufigkeiten von kausalem “denn” und “da” geschätzt nach manueller Durchsicht eines repräsentativen Anteils der Daten.

Twitter Terms of Service

- Suchfunktion Twitter Search liefert unvollständige Ergebnisse
- Twitter-Stream-Zugang ist ratenlimitiert
  - Aber für Deutsch kein Problem

- **Keine Weitergabe von aggregierten Tweets (=Korpus) erlaubt**
- Korpusweitergabe nur über Tweet-IDs möglich; einzelne Tweets müssen dann zeitaufwändig wieder gecrawlt werden, z.B. mit `https://github.com/lintool/twitter-tools`

- Löschung von Tweets und/oder Accounts: 21,2% des Tweets2011-Korpus verschwanden in den ersten 9 Monaten
- Anonymisierung von Tweets in Papieren
  - @-Tags entfernen
  - Trotzdem auffindbar