



Computerlinguistische Analyse von Twitterdaten

Gastvortrag im Seminar "Soziale Bewegungen im Internet"

FU Berlin

Tatjana Scheffler, Universität Potsdam

tatjana.scheffler@uni-potsdam.de

Sommersemester 2014: Universität Konstanz

15.5.2014

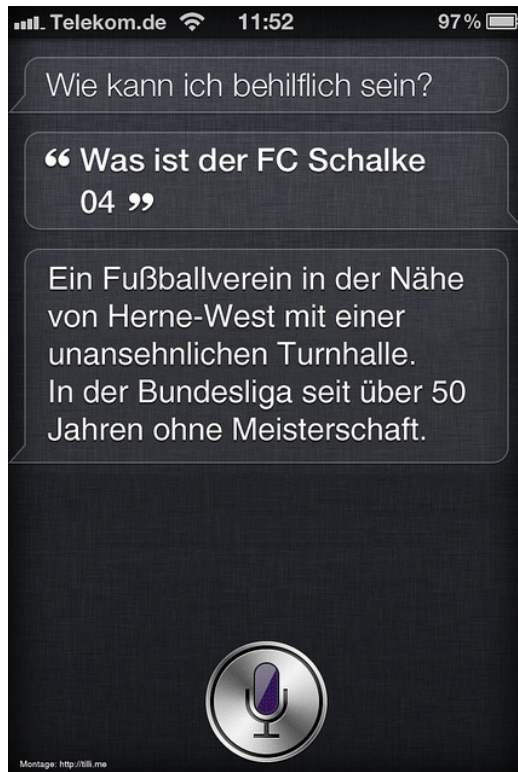
Plan

- ▣ Twitter und Computerlinguistik
- ▣ Korpuserstellung
- ▣ Analysebeispiele
- ▣ Tools und Methoden

Einführung

Twitter und Computerlinguistik

Computerlinguistik



- Analyse und Verarbeitung natürlicher Sprache mit dem Computer
- Analyse:
 - formale Repräsentationen
 - *Grammatik*
 - Korpuslinguistik
- Verarbeitung:
 - Anwendungen
 - Mensch-Maschine-Interaktion

Warum Twitter?

Für Linguisten/Computerlinguisten:

- sehr große Datenmenge (noch wachsend)
- in maschinenlesbarer Form im Netz
- aktuelle Themen
- viele Metadaten
- Spontansprache aus verschiedenen Genres
- spezieller Stil (zwischen geschriebener und gesprochener Sprache)

Praxis: Social Media Monitoring

- *Präsenzanalyse*: Statistische Analyse, die die Präsenz eines Zielkonzeptes im Web/Social Media angibt
- *Trendanalyse*: Was entsteht gerade?
- *Tonalitätsanalyse*: Meinungsbild der Zielgruppe
- *Buzz-Analyse*: Involvement einer Zielgruppe zu einem bestimmten Thema
- *Profiling*: Erkenne Meinungsführer und Multiplikatoren
- *Quellenanalyse*: Bedeutende Orte im Netz

Außerdem...

- ▣ Soziolinguistik
- ▣ Korpuslinguistik
- ▣ Diskursanalyse
- ▣ Twitter als empirische Datenquelle

Probleme bei der Analyse von Twitterdaten

- Bisherige Studien fast ausschließlich auf englischen Daten
- Twitter-Terms of Service verbieten viele forschungsrelevante Verwendungen der Daten
- Suchfunktion Twitter Search liefert unvollständige Ergebnisse
- Twitter-Stream-Zugang ist ratenlimitiert
 - Aber für Deutsch meist kein Problem
- <http://www.buzzfeed.com/nostrich/how-twitter-gets-in-the-way-of-research>

Twitter

- <http://www.twitter.com>
- Kurznachrichtendienst
- 140 Zeichen
- Follower-Friend-Beziehungen zwischen Nutzern
- Timeline aggregiert alle Nachrichten der Friends in Echtzeit
- @-Replies, Retweet-Relation, #Tag Themen
- Abrufen über Twitter API (JSON-Format)



Twitterdaten – Beispiel

- Leicht Vereinfachte JSON-Darstellung eines Tweets
- Attribut-Value Matrix
- (4 Folien)

```
$json (  
| text = "Cro: sehr, sehr dope! #XmasJam"  
| source = "Twitter for iPhone"  
| retweeted = FALSE  
| favorited = FALSE  
| retweet_count = 0  
| entities (  
| | user\_mentions => Array (0)  
| | (  
| | hashtags => Array (1)  
| | (  
| | | \[0\] (  
| | | | text = "XmasJam"  
| | | | indices => Array (2)  
| | | | (  
| | | | | \[0\] = 22  
| | | | | \[1\] = 30  
| | | | )  
| | | )  
| | )  
| | urls => Array (0)  
| | (  
| )  
)
```

```
|  place (
|  |   country = "Germany"
|  |   place_type = "city"
|  |   country_code = "DE"
|  |   name = "Stuttgart"
|  |   full_name = "Stuttgart, Stuttgart"
|  |   url = "http://api.twitter.com/1/geo/id/e385d4d639c6a423.json"
|  |   id = "e385d4d639c6a423"
|  |   bounding_box (
|  |   |   coordinates => Array (1) (
|  |   |   |   ['0'] => Array (4) (
|  |   |   |   |   ['0'] => Array (2) (
|  |   |   |   |   |   ['0'] = 9.038755
|  |   |   |   |   |   ['1'] = 48.692343 )
|  |   |   |   |   ['1'] => Array (2) (
|  |   |   |   |   |   ['0'] = 9.315466
|  |   |   |   |   |   ['1'] = 48.692343 )
|  |   |   |   |   ['2'] => Array (2) (
|  |   |   |   |   |   ['0'] = 9.315466
|  |   |   |   |   |   ['1'] = 48.866225 )
|  |   |   |   |   ['3'] => Array (2) (
|  |   |   |   |   |   ['0'] = 9.038755
|  |   |   |   |   |   ['1'] = 48.866225 ) ) )
|  |   |   type = "Polygon" )
|  |   attributes ( )
|  )
```

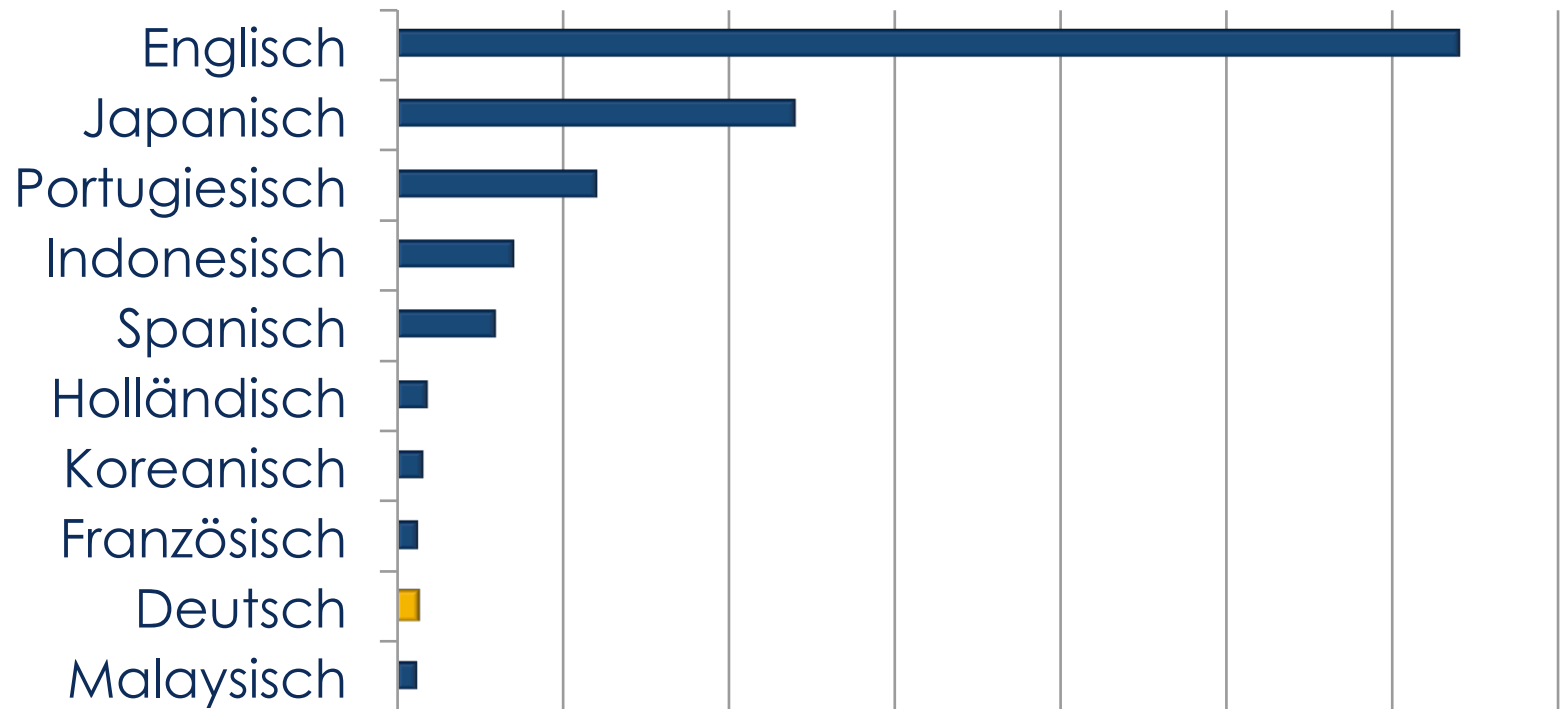
```
| user (  
| | friends_count = 1983  
| | follow_request_sent = NULL  
| | profile_sidebar_fill_color = "dbeefd"  
| | profile_background_image_url_https = "https://si0.twimg.com/...0210.jpg"  
| | profile_image_url = "http://a3.twimg.com/.../twitter_normal.gif"  
| | profile_background_color = "f1f9ff"  
| | url = "http://christianfleschhut.de/"  
| | id = 1182351  
| | is_translator = TRUE  
| | screen_name = "cfleschhut"  
| | lang = "en"  
| | location = "Karlsruhe, Germany"  
| | followers_count = 1628  
| | statuses_count = 3882  
| | name = "Christian Fleschhut"  
| | description = "93 â til"  
| | favourites_count = 166  
| | profile_background_tile = FALSE  
| | listed_count = 54  
| | created_at = "Wed Mar 14 21:15:22 +0000 2007"  
| | utc_offset = 3600  
| | verified = FALSE  
| | show_all_inline_media = TRUE  
| | time_zone = "Berlin"  
| | geo_enabled = TRUE  
| )
```

```
| truncated = FALSE
| in_reply_to_status_id_str = NULL
| created_at = "Thu Dec 22 21:22:36 +0000 2011"
| in_reply_to_user_id = NULL
| id = 149963070435893248
| in_reply_to_status_id = NULL
| geo (
| | coordinates => Array (2) (
| | | ['0'] = 48.78509331
| | | ['1'] = 9.18866308
| | )
| | type = "Point"
| )
| in_reply_to_user_id_str = NULL
| id_str = "149963070435893248"
| in_reply_to_screen_name = NULL
| )
```

Erstellung eines deutschen Twitterkorpus

Probleme, Vorgehensweise

Sprache auf Twitter

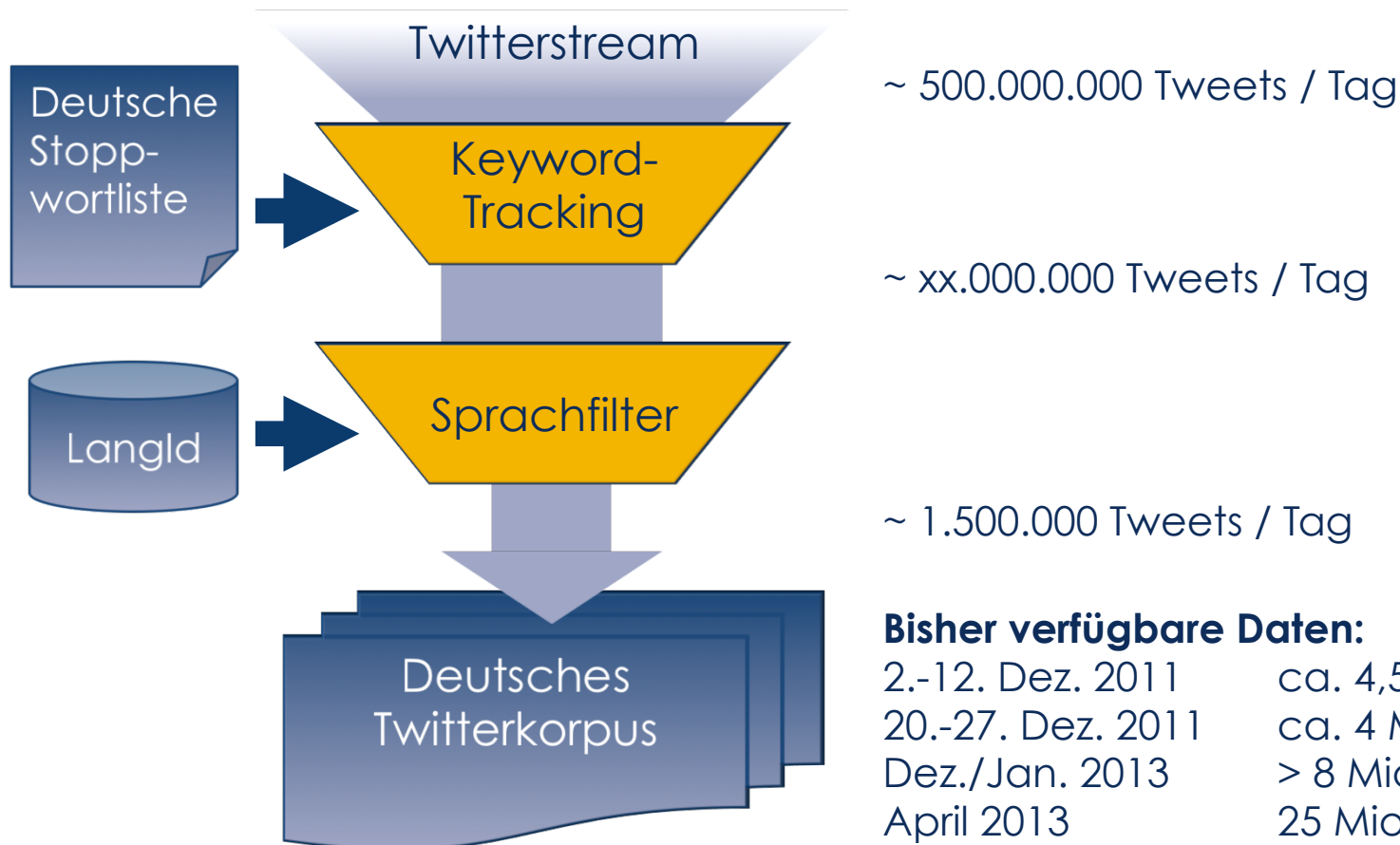


Quelle: Hong, Lichan, Convertino, Gregorio, and Chi, Ed. "Language Matters In Twitter: A Large Scale Study" International AAAI Conference on Weblogs and Social Media (2011)

Twitter-API zur Korpuserstellung

- Search API oder Streaming API
- Search API: Suchworte, ca. 7 Tage in die Vergangenheit
- Streaming API:
 - Echtzeitstream der entstehenden Tweets
 - Quotenlimitierung
 - Viele nicht-deutsche Tweets
 - Filter
 - Geolokation (location) < 2% der dt. Tweets
 - bis zu 5000 User-Ids (follow)
 - bis zu 400 Stichwörter (track)

Korpuserstellung



Tools: Twitterstream 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:
`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 als Korpus

- ▣ Enthält spezielle Tokens (Emoticons, URLs, # Hashtags)
- ▣ Umgangssprache, Slang und Dialekte
- ▣ **Vorverarbeitung ist wichtig:**
 - ▣ Normalisierung (Umlaute, Prolongationen, Tippfehler?)
 - ▣ Behandlung von Spezialtokens (@Handles, #Tags)
 - ▣ Tokenisierung
 - ▣ Satzgrenzenbestimmung

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

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

Twitter Terms of Service – Probleme

- **Keine Weitergabe von aggregierten Tweets (=Korpus) erlaubt**
- Korpusweitergabe nur über Tweet-IDs möglich; einzelne Tweets müssen 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
 - @Handles entfernen
 - Trotzdem auffindbar

Themen auf Twitter finden

Beispiel 1

Signifikante Wörter

- Analyse von Social Media-Daten zu einem Thema
- Einfachste Repräsentation der Textbedeutung: "Bag of Words"-Modell

Ziel 1: Auswahl relevanter Texte
für ein Thema

Ziel 2: Charakterisierung von Subkorpora

Ziel 3: Themen finden



Keywords (1/2)

- Chi-squared-Test zum Vergleich von Subkorpora
- Social Media vs. traditionelle Medien
- berechnet, ob Differenz zwischen erwarteter Frequenz E und beobachteter Frequenz O signifikant ist

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Keywords (2/2)

- Gegeben eine Startmenge (ein Seed-Korpus)
 - finde neues signifikant häufiges Vokabular
 - -> Vergrößerung des Korpus

- Tag1 vs. Tag2

- Trends verfolgen?

- Echtzeit?



Themen finden

- große Datenmenge automatisch clustern
- Reduzierung von Texten auf “Bag of Words”
- Säuberung
 - Entfernung von Füllwörtern
 - Lemmatisierung
- Jeder Text ist ein Vektor von Worthäufigkeiten
- Gewichtung der Wörter mit TF-IDF
(term frequency*inverse document frequency)

Ergebnis von k-means Clustering auf online Nachrichten

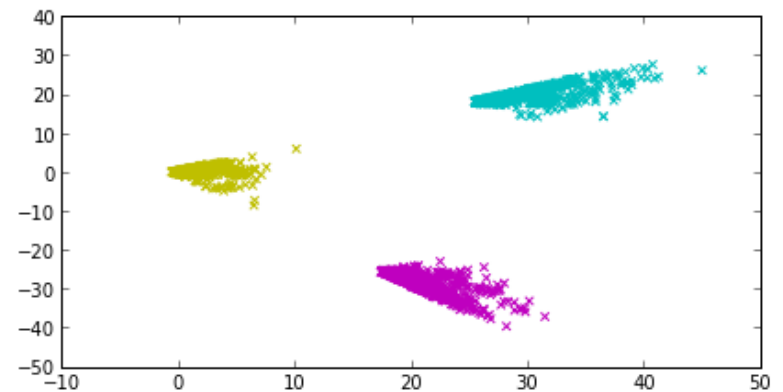
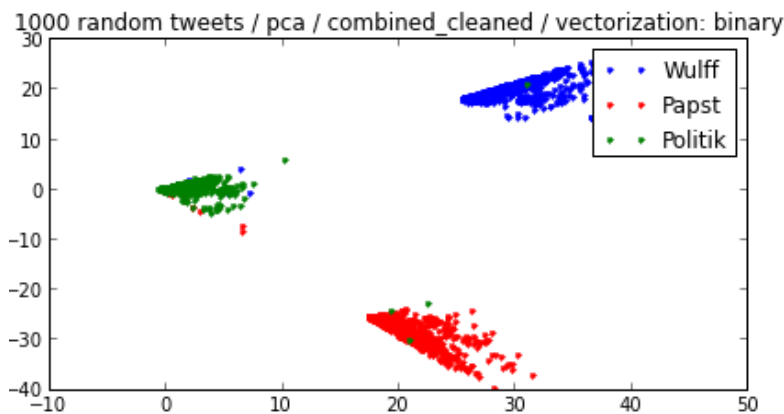
C	Docs	Thema (manuell)	Labels (automatisch)
1	11476	?	wulff, washington, brows, partei, putin, obama, mord, grun, militar
2	2044	Syrien	assad, obama, Syrien, damaskus, chemiewaff, regim, militar, syrischen, putin, Russland
3	1482	Bundestagswahl	steinbruck, Peer_Steinbrück, merkel, kanzlerin, wahlkampf, oradour, partei, Angela_Merkel, grun, SPD
4	1214	Bayern/Wahl	FDP, grun, csu, afd, seehofer, ude, Bayern, union, landtagswahl, bundestagswahl
5	251	Pädophilie-Skandal	trittin, grun, padophil, sexuell, gotting, franz_walter, jurg, padophil, agil, kommunalwahlprogramm
6	119	Papst/kath. Kirche	kirch, bischof, franziskus, van, kathol, papst, priest, kardinal, vatikan, Rom
7	110	Hessen/Wahl	bouffier, gumbel, schaf, hessen, landtagswahl, volker_bouffier, thorsten_schäfer_gümbel, grun, rot, landtag
8	62	NSU-Prozess	zschap, beate_zschäpe, anwalt, oberlandesgericht, richt, verteid, mord, hauptangeklagt, senat, angeklagt
9	53	Schulden/ Griechenland	griechenland, schaubl, schuldenschnitt, Athen, esm, griechenlands, troika, griechischen, milliard, finanzminist
10	33	Eurokrise	rezession, schatt, bruttoinlandsprodukt, defizit, arbeitslosenquot, konjunktur, zypern, italien, überschuss, portugal

Dargestellt sind die zehn größten Cluster; Stichprobe aller Politik-Feeds vom 26.08.–22.09.2013, n=16.953, k=19.

Elisabeth Günther & Thorsten Quandt; Jahrestagung der FG
Computervermittelte Kommunikation der DGPK 2013 Wien,
7.-9.November 2013

Topic Detection in Social Media

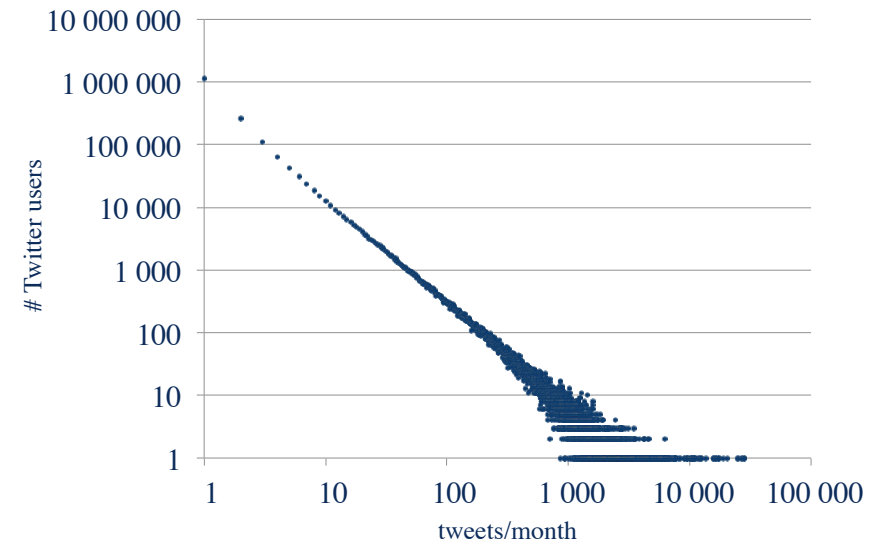
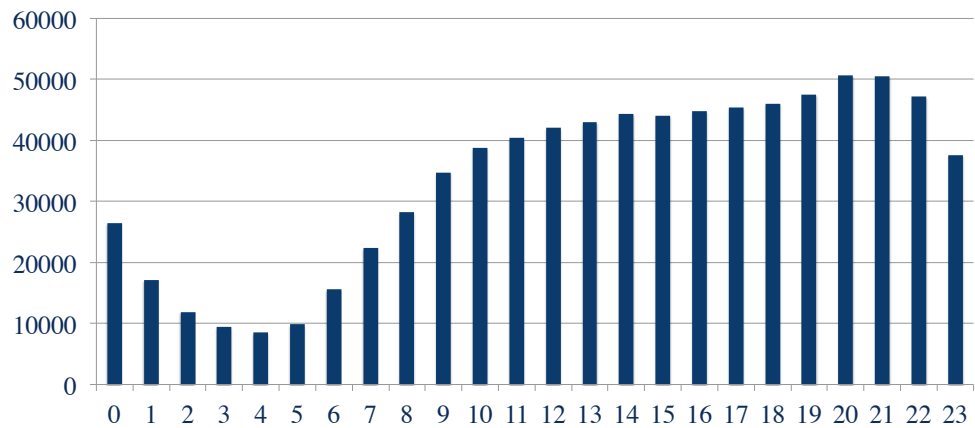
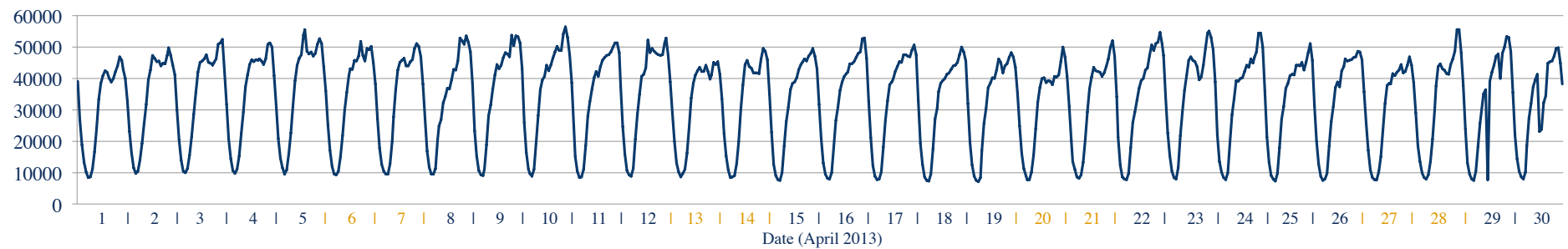
- Probleme!
 - extrem kurze Texte
 - nicht standardkonforme Sprache
- Clustering von freien Tweets daher zur Zeit schlecht



Diskurse

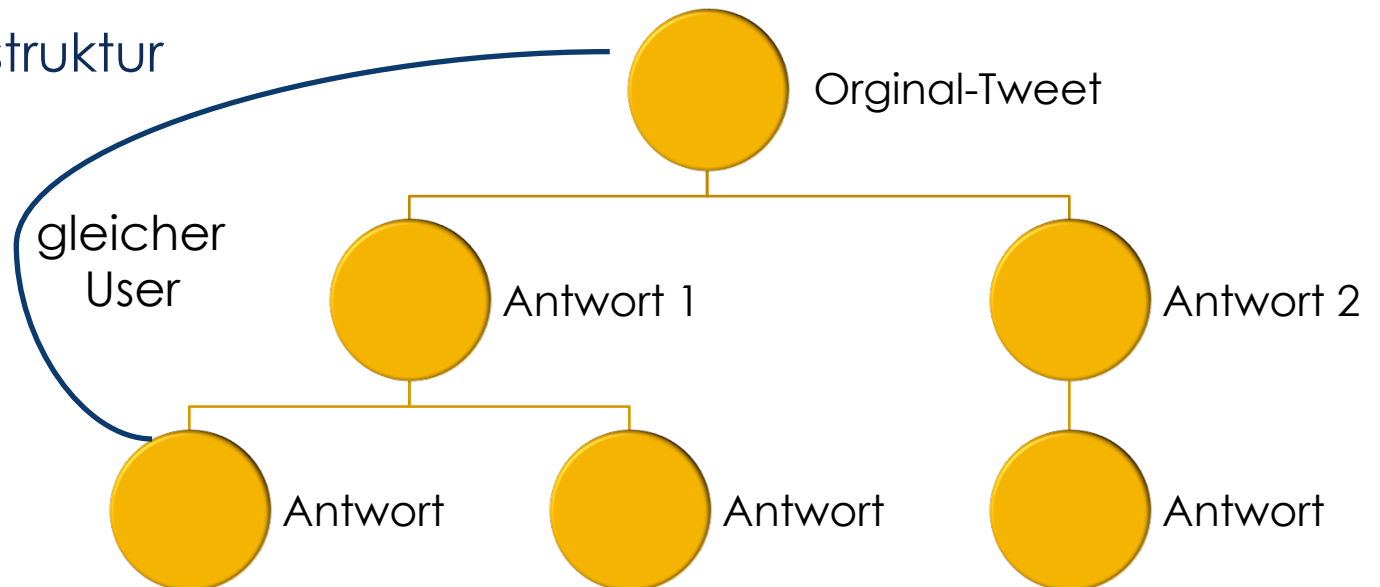
Beispiel 2

Deutsche Twitterdaten



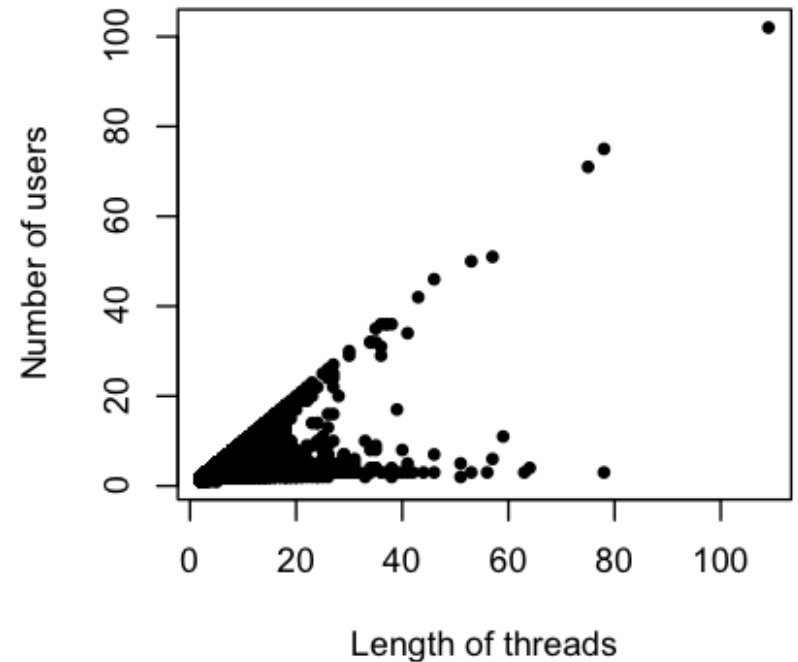
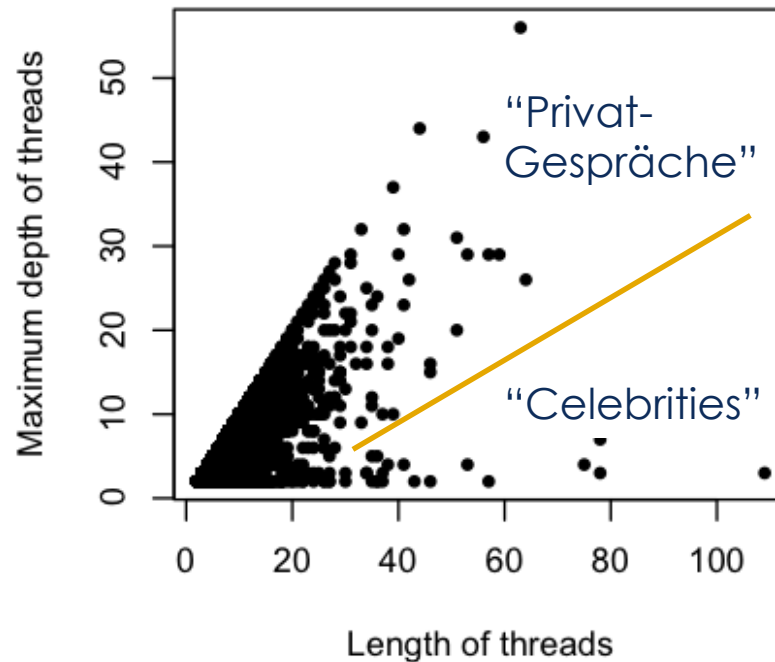
Diskurse auf Twitter

- Reply-to-Funktion strukturiert Tweets in *Diskurse*
- 21.2% der dt. Tweets sind Antworten
- Baumstruktur



Diskursstruktur

■ Verschiedene Typen von Diskursen



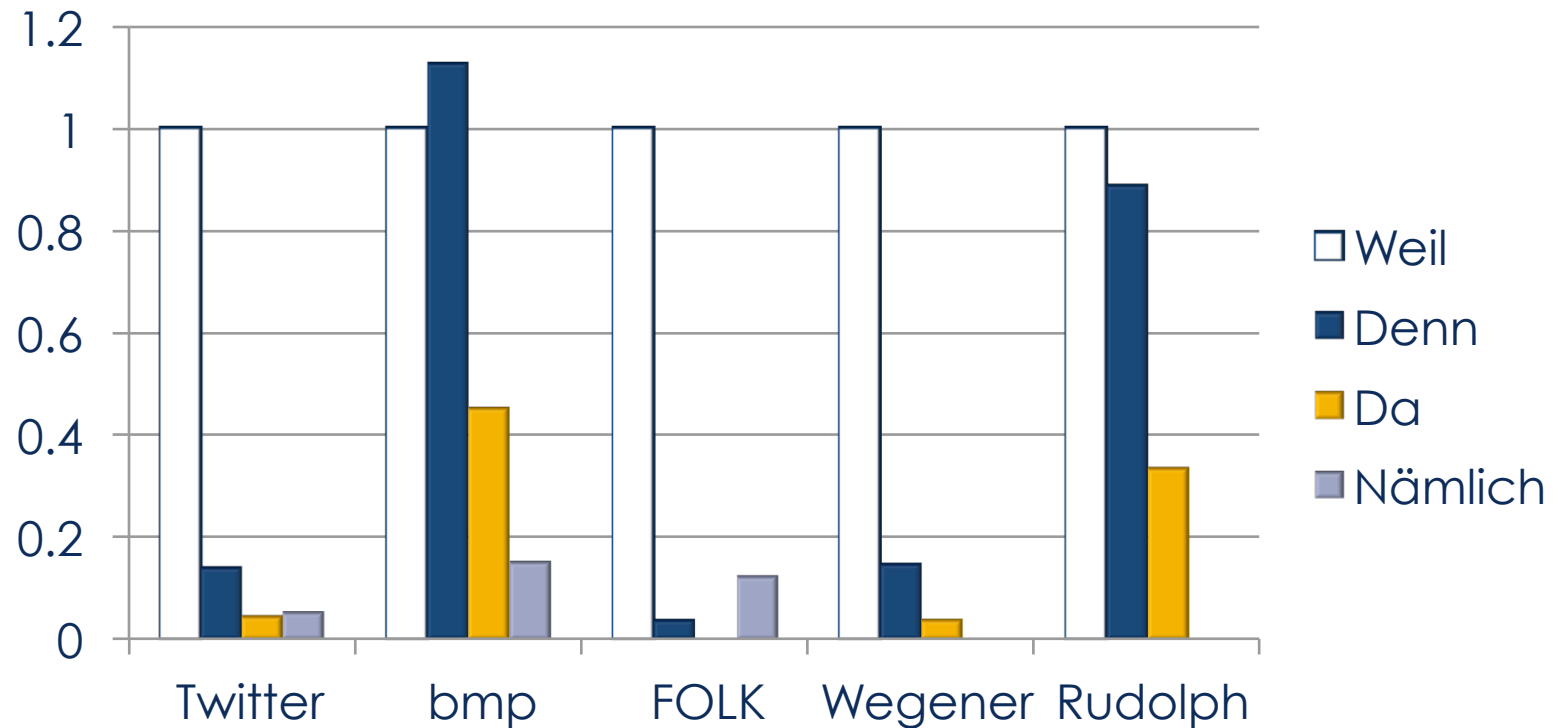
Deliberative Diskurse

- Habermas (1981): Rechtfertigungen für Behauptungen
- Kausalkonnektoren sind häufig auf Twitter:
 - 1.7% aller Tweets
 - 2.6% aller Antworten
- “gesprochener”/informeller Begründungsstil

Wir waren bei mir weil wir hatten Ausfall

Wer leiht Wulff eigentlich das Geld für seine
Anwälte? Ganz billig ist die Veranstaltung nämlich
nicht.

Twitter-Stil: Kausalkonnektoren



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)

Sentimentanalyse

Beispiel 3

Sentimentanalyse

WTF? Ich habe Naturstrom und soll **jetzt Kohle- und Atomstrom mitfinanzieren**? Was für ein Unsinn. WAS FÜR EIN UNSINN!

- Finden von subjektiven Äußerungen
 - Meinung
 - Ziel der Meinung
 - Quelle der Meinung (Meinungsträger)
- Korpusannotation für Trainingsdaten
- Maschinelle Lernverfahren (CRF)

Tools und Praktisches


... für Ihre weitere Arbeit

Sentimentanalyse (1)



http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/

Sentimentanalyse (2)



Multi-Perspective
Question Answering

Main
MPQA Home

Corpora
News, debates, etc.

Lexicons
Subj. clues, etc.

Annotation
GATE, MPQA, gfbf

OpinionFinder
Subjectivity detector

OpinionFinder

[Version 1.x](#)

[Version 2.x](#)

- o **OpinionFinder System**

OpinionFinder is a system that processes documents and automatically identifies subjective sentences as well as various aspects of subjectivity within sentences, including agents who are sources of opinion, direct subjective expressions and speech events, and sentiment expressions. OpinionFinder was developed by researchers at the University of Pittsburgh, Cornell University, and the University of Utah. In addition to OpinionFinder, we are also releasing the automatic annotations produced by running OpinionFinder on a subset of the Penn Treebank. To go to the OpinionFinder download page click [here](#).

contact: mpqa.project@gmail.com

[\[nlp\]](#) [\[cs\]](#) [\[pitt\]](#)

<http://mpqa.cs.pitt.edu/opinionfinder/>

Fragen?

tatjana.scheffler@uni-potsdam.de

Projekt: Analyse von Diskursen in Social Media,
funded by BMBF, # 01UG1232A

Web: <http://www.social-media-analytics.org/>

Bildreferenzen

Siri - by-nc-sa Henning Tillmann <https://www.flickr.com/photos/henningtillmann/6246291025/>