Dialoge auf Twitter

Seminar, AM7

Universität Potsdam

Tatjana Scheffler

tatjana.scheffler@uni-potsdam.de

14.10.2015

Fragestellungen

- Welche besonderen sprachlichen Phänomene gibt es in Dialogen?
- Wie koordinieren Dialogteilnehmer ihre Beiträge?
- Wie kann die Struktur von Dialogen beschrieben werden?
- Welche Anforderungen müssen Modelle erfüllen?
- Wie können die vorhandenen Modelle und Ansätze auf Twitterdaten übertragen werden?

Twitter

… von “Statusmeldungen” zu Konversationen

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

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)

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

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 ( )
|   )
```

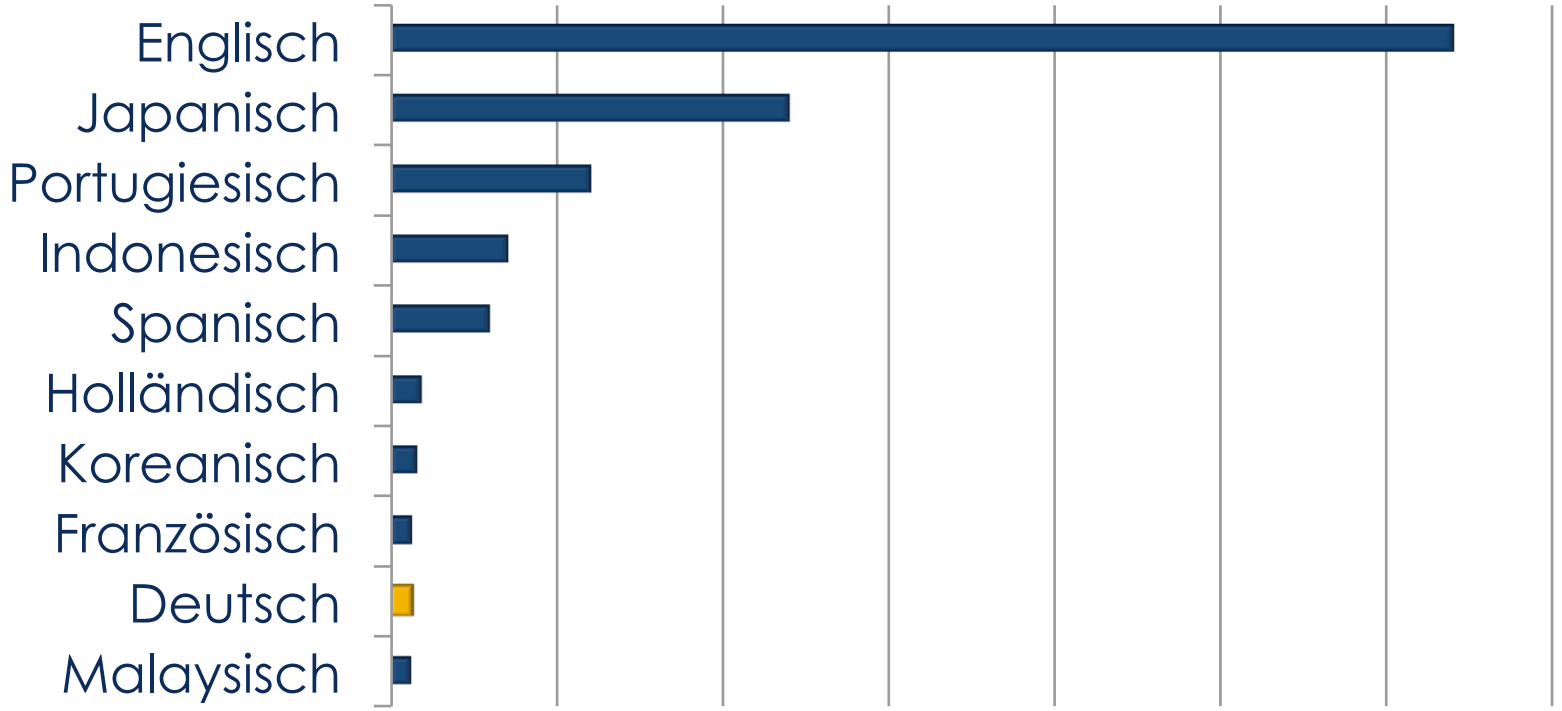
```
|   user (
|     |   friends_count =  1983
|     |   follow_request_sent =  NULL
|     |   profile_sidebar_fill_color =  "dbeefd"
|     |   profile_background_image_url_https = "https://si0.twimg.com/...0210.jpg"
|     |   profile_image_url =  "http://a3.twimg.com/…/twitter_normal.gif"
|     |   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 =  14996307043589324 8
|    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 =  "14996307043589248"
|    in_reply_to_screen_name =  NULL
)
```

Erstellung eines deutschen Twitterkorpus

Probleme, Vorgehensweise

Sprache auf Twitter

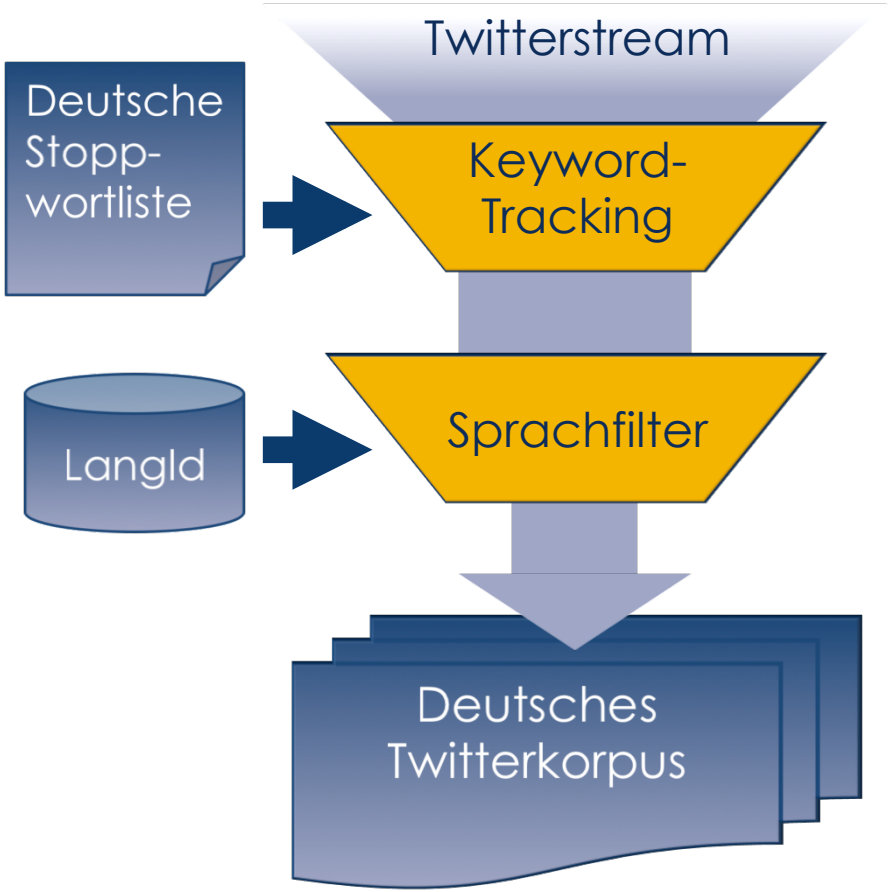


Quelle: Hong, Lichan, Convertino, Gregorio, and Chi, Ed. "Language Matters In Twitter: A Large Scale Study" International AAAI Conference on Weblogs and Social Media (2011)

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



~ 500.000.000 Tweets / Tag

~ xx.000.000 Tweets / Tag

~ 1.500.000 Tweets / Tag

Bisher verfügbare Daten:

2.-12. Dez. 2011	ca. 4,5 Mio. Tweets
20.-27. Dez. 2011	ca. 4 Mio. Tweets
Dez./Jan. 2013	> 8 Mio. Tweets
April 2013	25 Mio. Tweets ↗

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
```

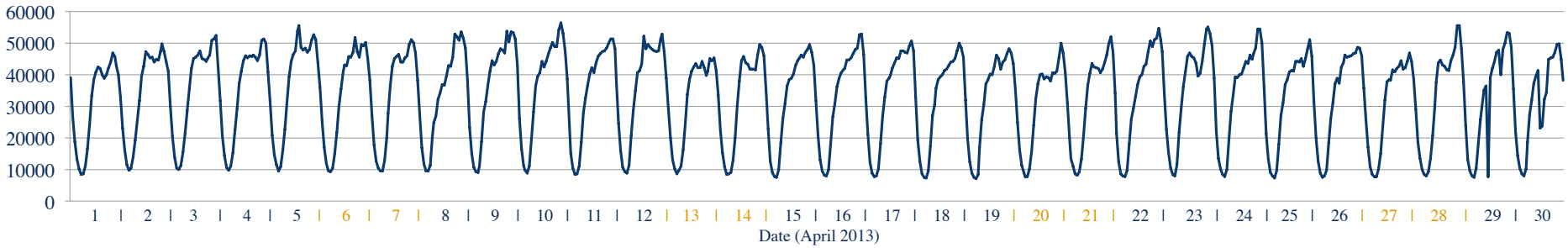
Twitter Terms of Service – Probleme

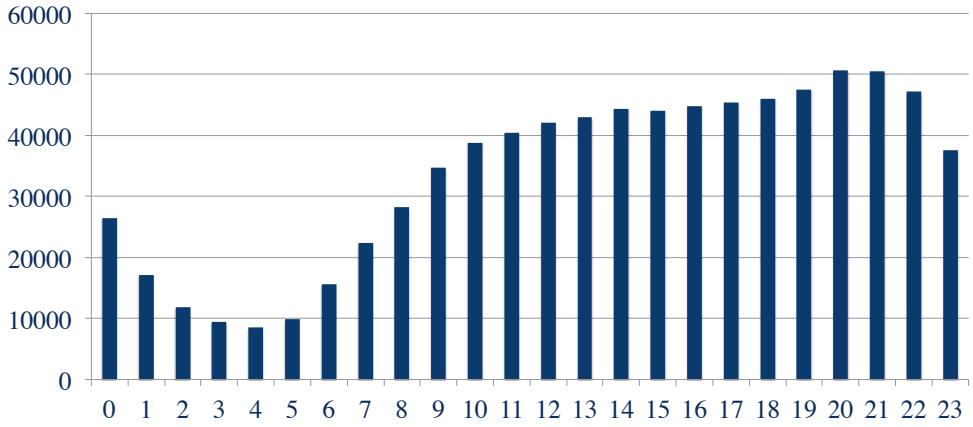
- **Keine Weitergabe von aggregierten Tweets (=Korpus) erlaubt**
- Korpusweitergabe nur über Tweet-IDs möglich; einzelne Tweets müssen zeitaufwändig wieder gecrawlt werden, z.B. mit `https://github.com/lintool/twitter-tools`
- Löschung von Tweets und/oder Accounts: 21,2% des Tweets2011-Korpus verschwanden in den ersten 9 Monaten
- Anonymisierung von Tweets in Papieren?
  - @Handles entfernen
  - Trotzdem auffindbar

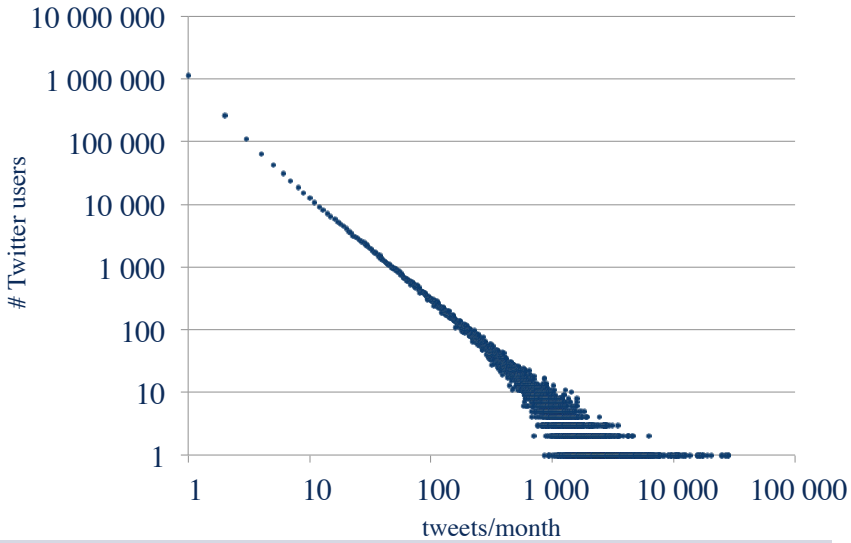
Diskurse

Konversationen auf Twitter

Deutsche Twitterdaten

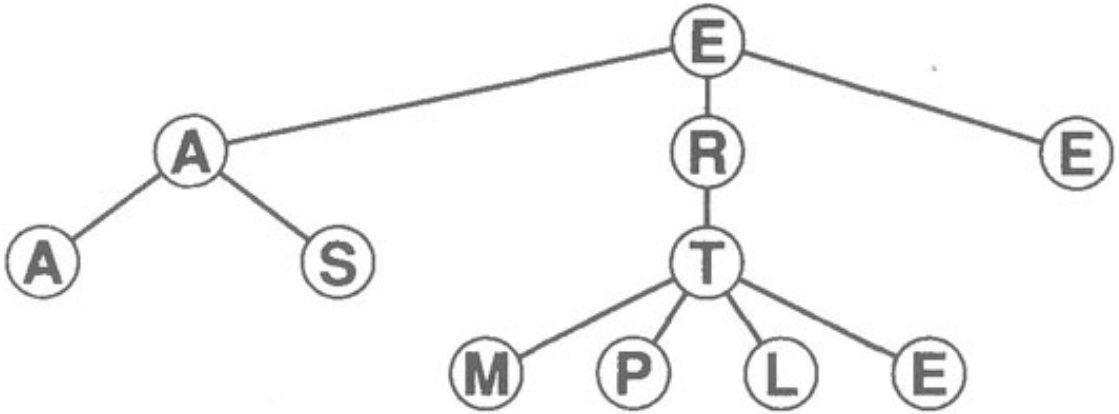






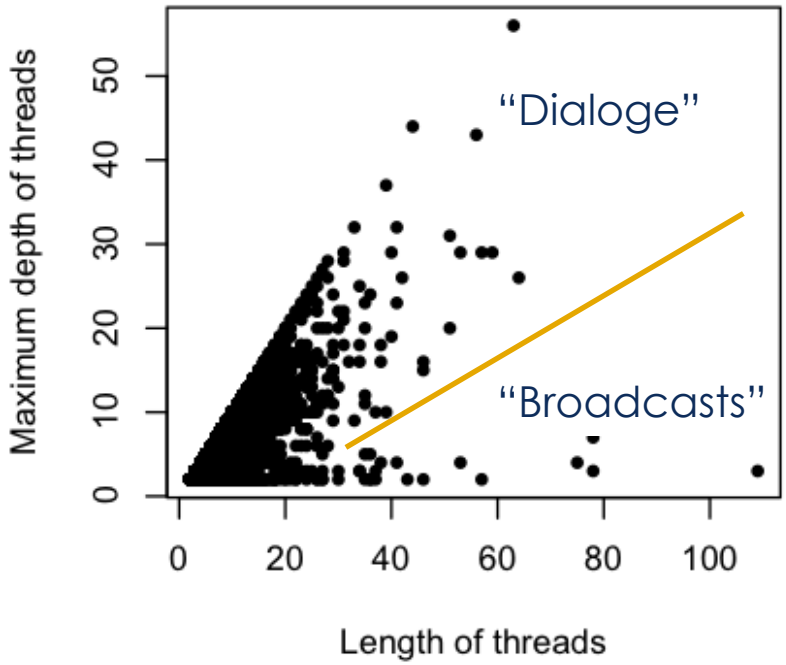
Konversationen auf Twitter

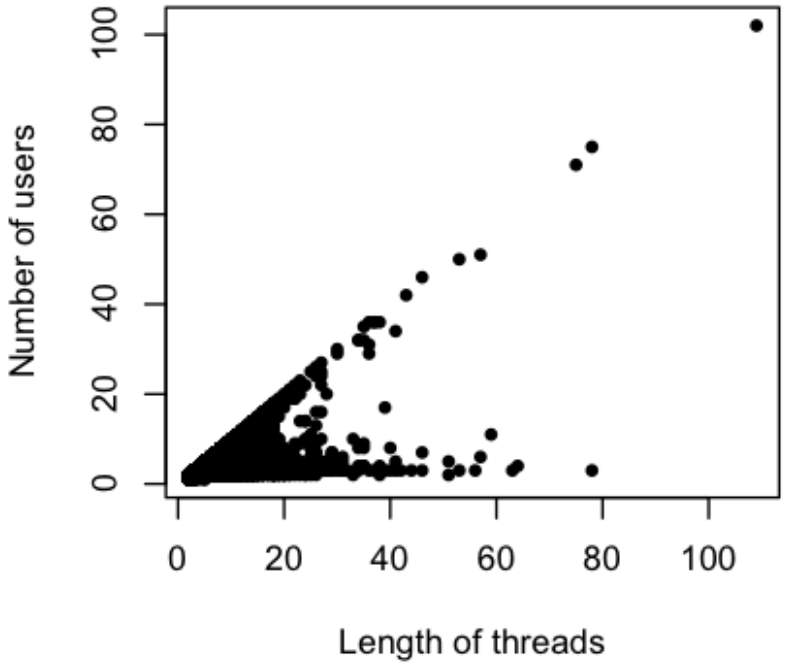
- Reply-to-Funktion strukturiert Tweets in *Diskurse*
- ~20-25% der dt. Tweets sind Antworten
- Baumstruktur



Diskursstruktur

- Verschiedene Typen von Diskursen



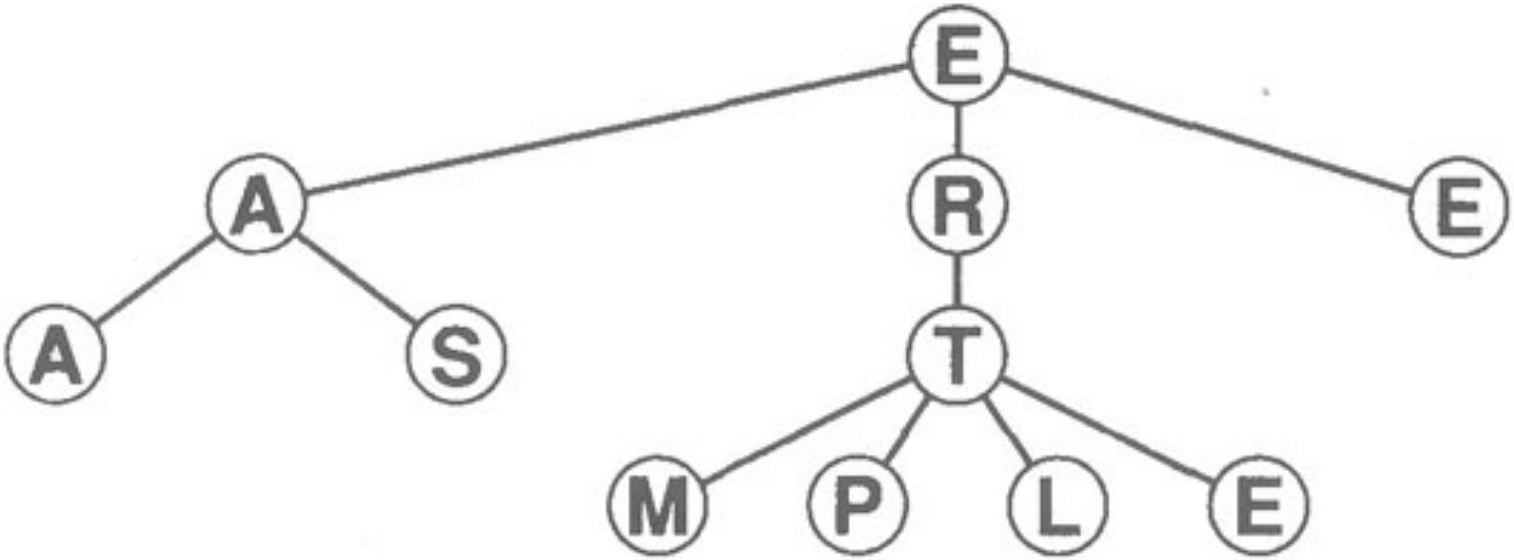


Twitter-Dialoge: Relationen

- @TheBug0815 @McGeiz Das sind dumme Behauptungen. Ich helfe gerne weiter, Dir zu erklären, wie eine Stromnetz funktioniert & was Zufallsstrom ist
- @McGeiz **doch**, es ist ein flexibler Park aus Speichern und Gaskraftwerken nötig, Kohle muss schnell weg @willlistock @Luegendetektor

- #Offshore-Ausbau: Warum schweigen Dauer-#EEG-Kritiker @Der_BDI @iw_koeln @insm @DICEHHU @RolandTichy @bdew_ev @igbce? http://t.co/WfZsirxMiC
- @UdoSieverding **weil** sie alle Interessen der großen Stromkonzerne vertreten @Der_BDI @iw_koeln @insm @DICEHHU @RolandTichy @bdew_ev @igbce

Twitter-Dialoge: Relationen



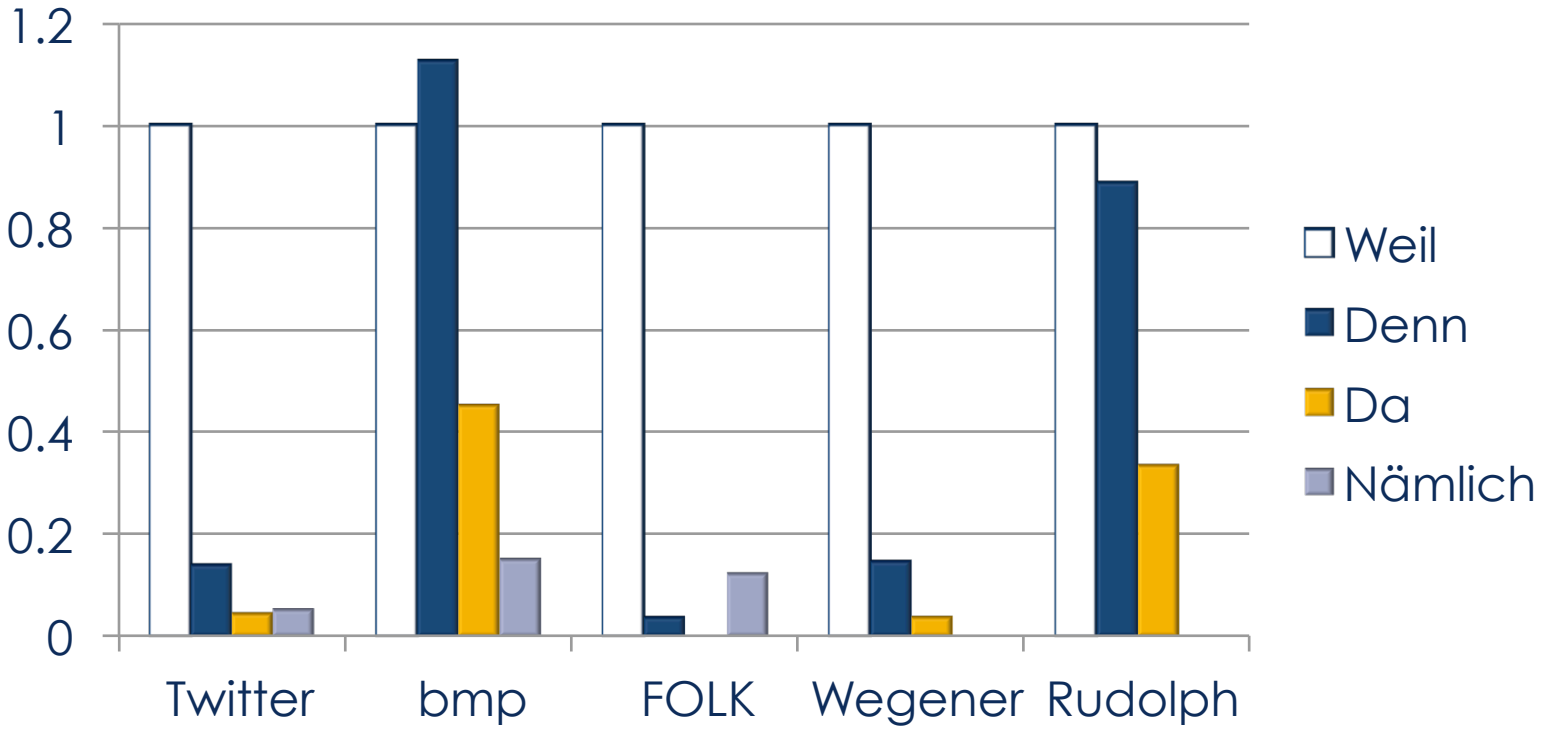
Konnektoren auf Twitter

- 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)

Dialoge

Struktur und Modellierung

Natürliche Sprache?

- Die Kündigung durch den bisherigen Sozialarbeiter habe den Gemeinderat aber zu einer Lageanalyse veranlasst, so Stadtpräsident Roger Hochreutener. **Weil** im nächsten Jahr die regionale Kinder- und Erwachsenenschutzbehörde (KES) in Bütschwil ihren Betrieb aufnehme, würden sich **nämlich** die Aufgaben der lokalen Sozialberatung reduzieren.

- Verschiedene Leserinnen und Leser haben sich an die Redaktion gewandt und ihre Verwunderung oder sogar Ärger ausgedrückt. **Weil nämlich** Gächter die Wahlempfehlung für sich selbst und drei seiner Parteikollegen mit einem alten Grenzwacht-Stempel verschickte, suggeriert er - bewusst oder unbewusst -, dass auch die Grenzwacht SVP-Kandidaten zur Wahl empfehle.

Natürliche Sprache?

- Ich weiß, wie die Straße heißt: Uniwiesitätsstraße. **Weil nämlich** da vorne die Uniwiese ist
- **Weil nämlich**: auch für Frau Bergmann ist was dabei.
- Ja genau, **weil nämlich** nur Männer keinen Salat mögen
- Da hast du was falsch verstanden...das lesen noch weniger, **weil nämlich** nur die, die uns beiden folgen;)

Diskursphänomene (Monolog)

Peter went to John's party.
He drank all the wine.

- Anaphern
- Kohärenzrelationen

- geschriebener Text
- ohne Fokus auf Kommunikation
- Sprache als Produkt, nicht Prozess

Dialog

- gemeinsamer Prozess von mind. 2 TeilnehmerInnen
- gleichberechtigte autonome Agenten
- Kommunikation hauptsächlich durch spontane gesprochene Sprache
- normalerweise in Face-2-Face-Situationen
- Kollaboration und Kooperation der Gesprächspartner

Rand/Negativbeispiele?

- weniger prototypische Dialogsituationen?

- keine Dialogsituation?

Problem 1: Analyseeinheit

> um it'll be there it'll get to Dansville at three a.m. and then you wanna do you take tho-- want to take those back to Elmira so engine E two with three boxcars will be back in Elmira at six a.m. is that what you want to do?

- Satz?

Problem 2: Disfluency

until you're at the le I mean at the right exit

- Selbstkorrektur
- (mind.) 2 Konversationsstränge:
  - Thema des Dialogs
  - Metakommunikation

Problem 3: Grounding

A: Wer kam zur Party?
B: Welche Party?

Problem 4: Sprachl. Handlungen

Weißt Du, wie spät es ist?

———

Es zieht!

- in Face-2-Face-Dialogen werden Handlungen auch nichtsprachlich durchgeführt

Seminarplan

Zurück zu Twitter + Dialog

- nicht gesprochen
- nicht Face-2-Face
- informell
- dialogisch (Konversationsstruktur)
- semi-spontan
- enthält viele spontansprachliche Phänomene

Themen (vorläufig)

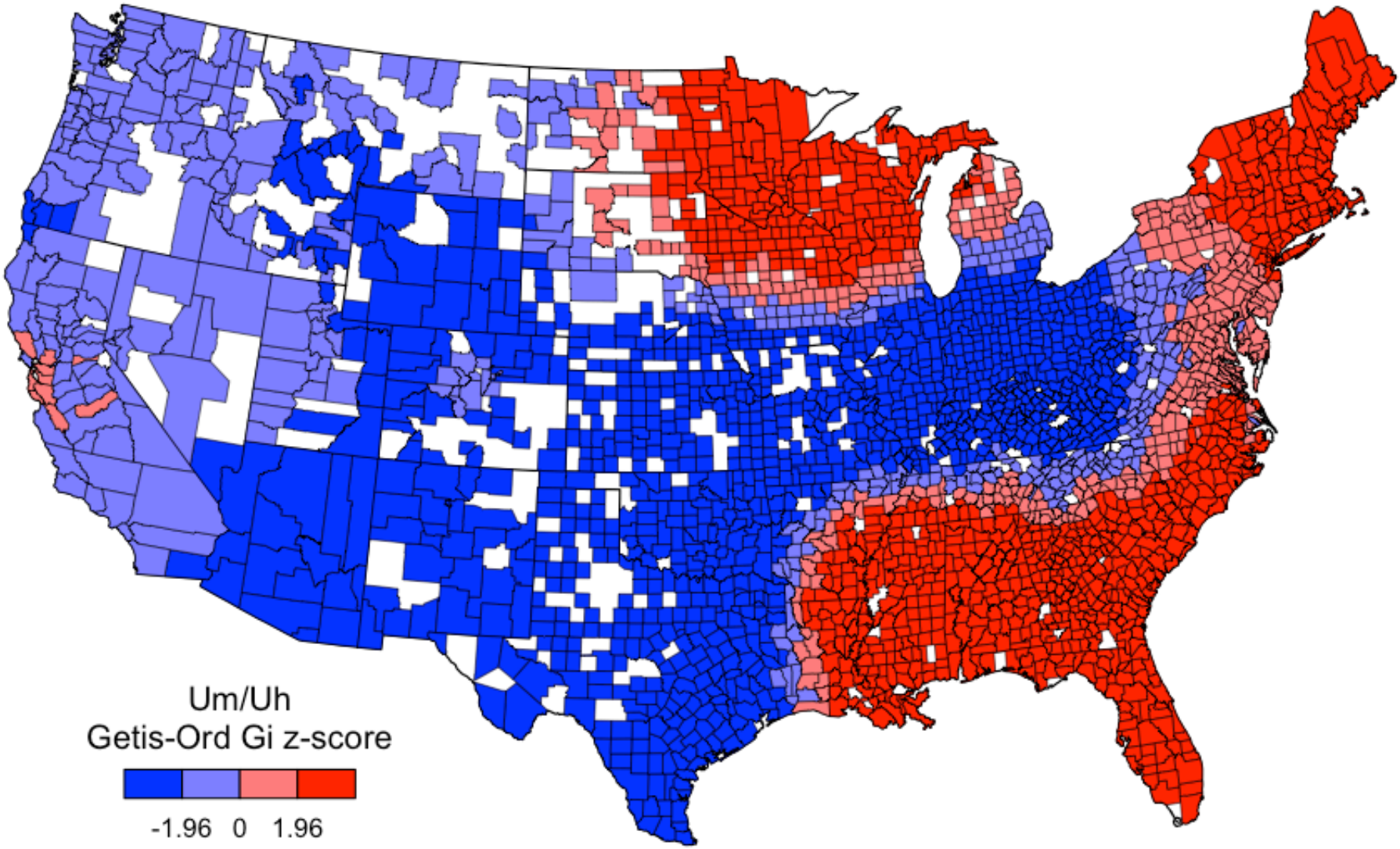
- Diskursstruktur
- Äußerungen
- Turn-Taking
- Grounding
- Dialogakte
- Anaphora/Referenz in Dialogen
- Dialogsysteme
- Ling. Phänomene: Fragen, Partikel
- Agreement/Disagreement, Stance
- Sarkasmus

Aufgabe!

1. Bei Twitter anmelden
2. 5 Leuten folgen, auch dem Kursaccount @twitling15
3. Suchen Sie je drei linguistische Merkmale/Phänomene, wo sich Twitterkonversationen und gesprochene Dialoge ähneln bzw. unterscheiden. Finden Sie Beispiele dafür!

DANKE

tatjana.scheffler@uni-potsdam.de


Um/Uh
Getis-Ord Gi z-score
-1.96 0 1.96

Bildreferenzen

- Dokument - By Silvestre Herrera (Author's website) [see page for license], via Wikimedia Commons
- um / uh – Jack Grieve, https://sites.google.com/site/jackgrieveaston/treesandtweets , August 18, 2014
- Sonstige Grafiken aus:

  Tatjana Scheffler. A German Twitter Snapshot. In: *Proceedings of LREC*, Reykjavik, Iceland. 2014.

  und von den Postern unter:

  http://www.ling.uni-potsdam.de/~scheffler/twitter/index.html