Telling Twitter apart
Eine Analyse zu Varietäten des Englischen auf Twitter

Netaya Lotze, Asmelash Teka Hadgu, Saskia Kersten

Twitter über Twitter

“connects you to the latest information”

“small bursts of information”

“you can share a lot with a little space”

(twitter.com/about 2013)

“connect with your friends”

“watch events unfold”

(twitter.com 2015)

Was ist Twitter?

- “publishing service”, im Sinne eines micro-blogs (Heil & Piskorski 2009: 2)

- “SMS of the Internet” (Crystal 2011: 36)

Aufbau der Studie

Small Scale Study

- 960 Tweets, manuelle Annotation
- Teilkorpora: GB, US, AUS

Large Scale Study

- 6,5 Mio. Tweets, automatische Annotation
- Teilkorpora: GB, US, AUS

Small Scale Study
Datenbasis

- Korpuserstellung im Sommer 2010 (GB & US) und 2013 (für AUS)
  - aus der public timeline von twitter.com
- 320 Tweets pro Varietät
  - 10 Tweets pro UserIn, 50% Userinnen, 50% User
- Tagging
  - PoS: automatisch (Penn TreeTagger 2.0), Nachbearbeitung von Hand
  - andere Analyseparameter: manuell

Varietäten auf Twitter

“typically understood to be based in territorial speech communities [....]
[and] questions about discourse, situated language use [...] tended to recede into the background”

(Mair 2013:254)

→ *Communities of Practice* (CoPs) als mögliche alternative Beschreibungsebene?

(s. Bamman, Eisenstein & Schnoebelen 2012)

Small Scale Study: Ergebnisse
Orthographie

- Von allen Wortformen sind 89,3% im GB-Korpus, 90,6% im AUS-Korpus und 93,4% im US-Korpus korrekt geschrieben
- Wenn z.B. Grapheme ausgelassen werden, scheint dies meist bewußt zu geschehen, um eine bestimmte Effekt zu erzielen:

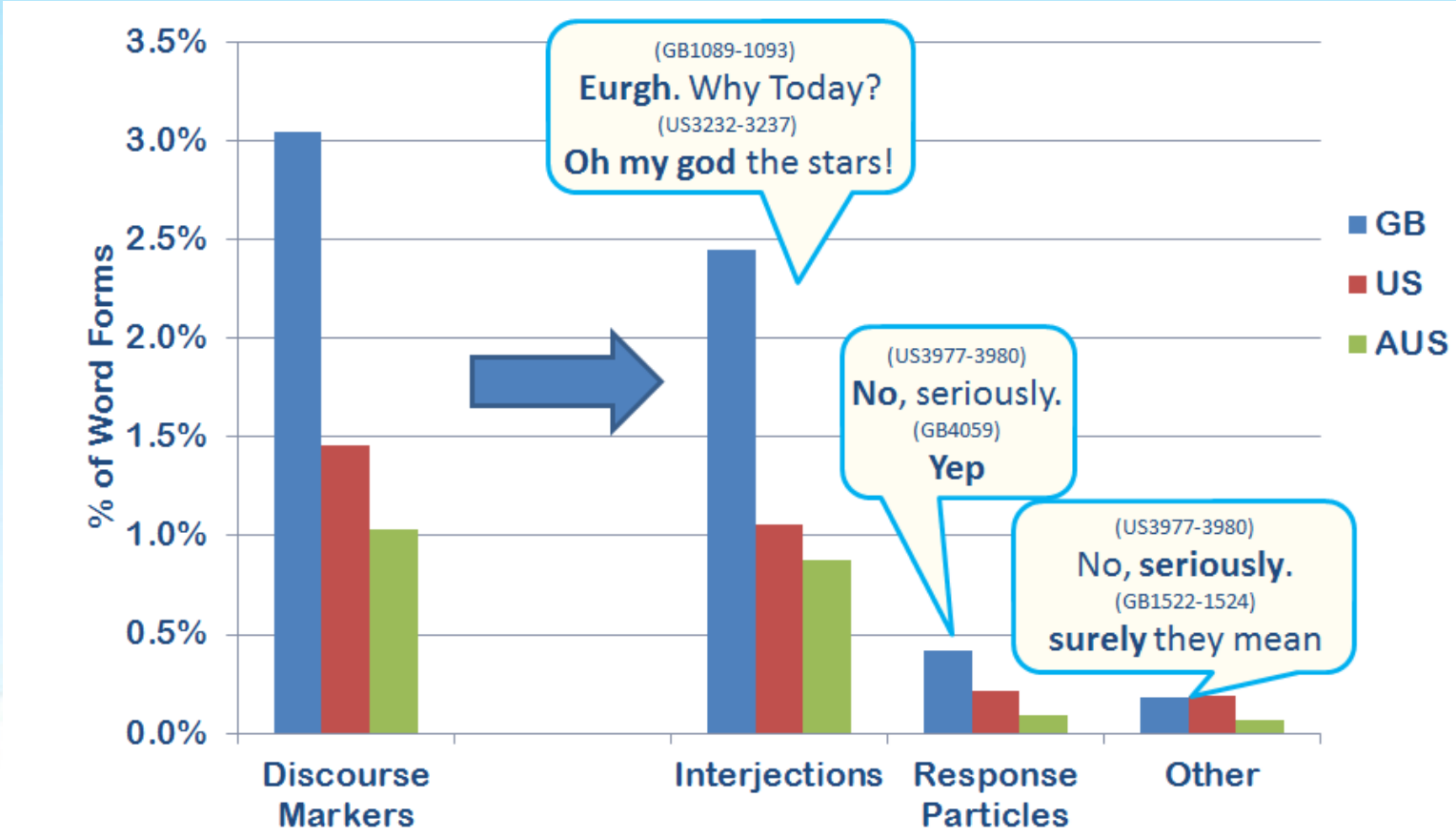
  (GB361-369) I’m not on ere to promote anything.

Small Scale Study: Ergebnisse

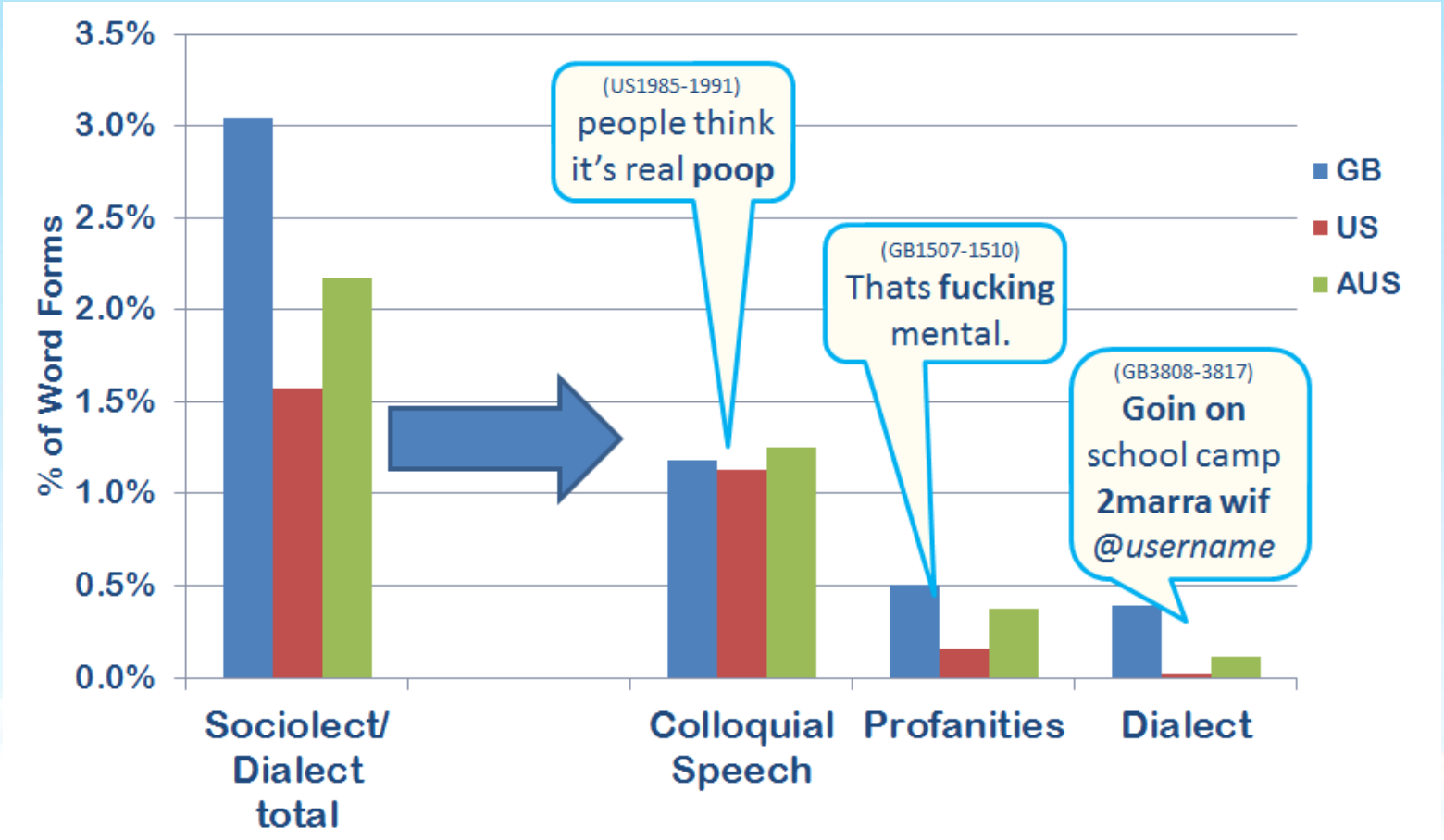
The Sad Apostrophe

Orthographie	GB	GB % WF	US	US % WF	AUS	AUS % WF
Auslassung des Apostrophs	44	1,01%	5	0,12%	18	0,39%

Diskursmarker



Lexis


3.5%
3.0%
2.5%
2.0%
1.5%
1.0%
0.5%
0.0%
% of Word Forms
(US1985-1991)
people think it's real poop
(GB1507-1510)
Thats fucking mental.
(GB3808-3817)
Goin on school camp 2marra wif @username
GB
US
AUS
Sociolect/ Dialect total
Colloquial Speech
Profanities
Dialect

Small Scale Study
Ergebnisse

- Unterschiede zw. Varietäten auf Twitter vor allem in Spektrum von Nähe und Distanz (Koch & Oesterreicher 1994)
- zahlreiche Belege für *Sprache der Nähe* in GB-Korpus (*konzeptionelle Mündlichkeit*, *Sprachökonomie*)

Gründe

- funktional (dialogische Sequenzen)
- medial (Verfügbarkeit von Autokorrektur)

Large Scale Study
“Telling Twitter Apart”

auf Basis der explorativen Studie wurden folgende Merkmale als mögliche Prädiktoren für das Britische ausgewählt:

- Auslassungen:
  - Apostroph (*dont*, *im*)
  - Graphem (*ere*)
  - Subjektpronomen (diary drop)
  - finites (Hilfs)verb
- Kleinschreibung von *I*
- Interjektionen
- Gebrauch von Schimpfwörtern
- Iteration von Satzzeichen

Dataset

- Crawl using the **Twitter Streaming API** tracking 6 cities
  - US - New York, Houston
  - UK - London, Birmingham
  - AUS - Sydney, Perth
- Bounding box obtained from **http://boundingbox.klokantech.com**
- Duration: 07 - 27 April 2014 inclusive
- Roughly 14M, 4M, 300K tweets for US, GB and AUS respectively

Pre-processing

1. **User Filtering**
    a. Baby names as proxy to filter users within their respective countries. Dataset scraped from open government data
    b. Keep users whose language is set to English
    c. Activity filter: a user who has at least 10 tweets at the time of crawl

1. **Tweet Filtering**
    a. Remove retweets and replies
    b. Remove tweets containing less than threshold of 2 tokens
    c. Keep tweets that are written in English

Features

I. **Linguistic features**
   using POS tagger specifically built for Twitter **Tweet NLP**
   A. Interjections: check if POS tagger has flagged "!" in a tweet
   B. Omission of apostrophes
      - nominals without apostrophe: if tag == 'L' & apostrophe not in token
      - contracted verbs: if tag == 'V' and token in list of contracted verbs
   A. Swear words: collected from **banned word list** and **front gate media terms to block**

I. **Token n-gram**
   Taking character n-grams of lengths 2 to 6

Classification

Problem definition: Given a tweet in English classify as AUS, GB, US

- Approach: supervised machine learning
- Using Linguistic and Statistical Features
- Training and testing set with stratified sampling: 5K tweets from each city using 5 fold cross-validation with **scikit-learn**
- Preliminary result:

Feature type	Algorithm	Precision	Recall	F1-score
Linguistic	Linear SVM	0.46	0.47	0.46
char n-grams	Naive Bayes	0.64	0.64	0.63

Next steps

- Training examples for non-classifiable English tweets (i.e., not AUS, GB or US) e.g., taking official sources such as news outlets

- Comparing **automatic feature learning** using deep learning against linguistic and statistical baselines

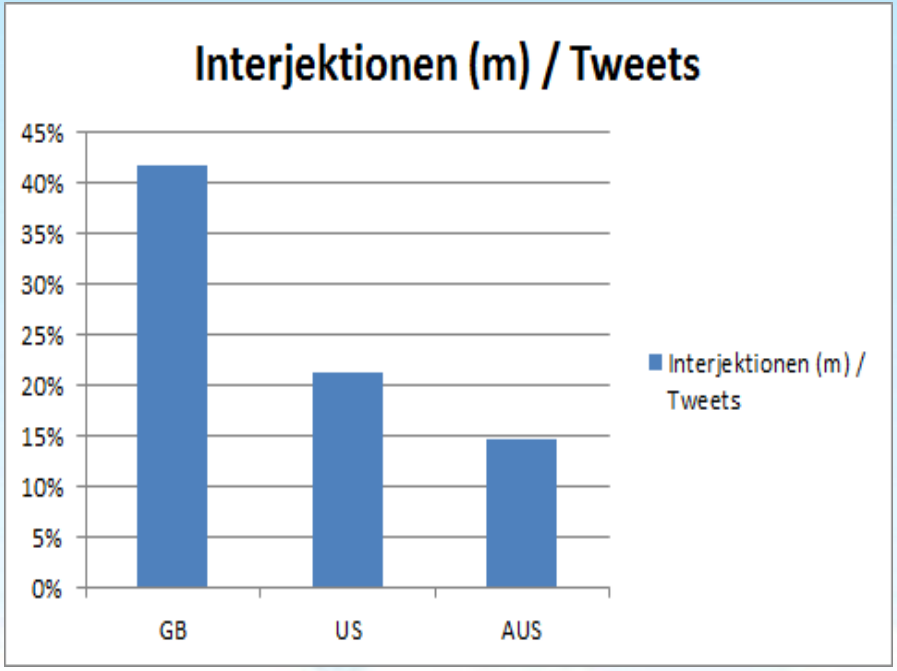
Small Scale vs. Large Scale Study
Vergleich der Ergebnisse

- Unterschiede im Nähe-Distanz-Spektrum beim größeren Sample geringer

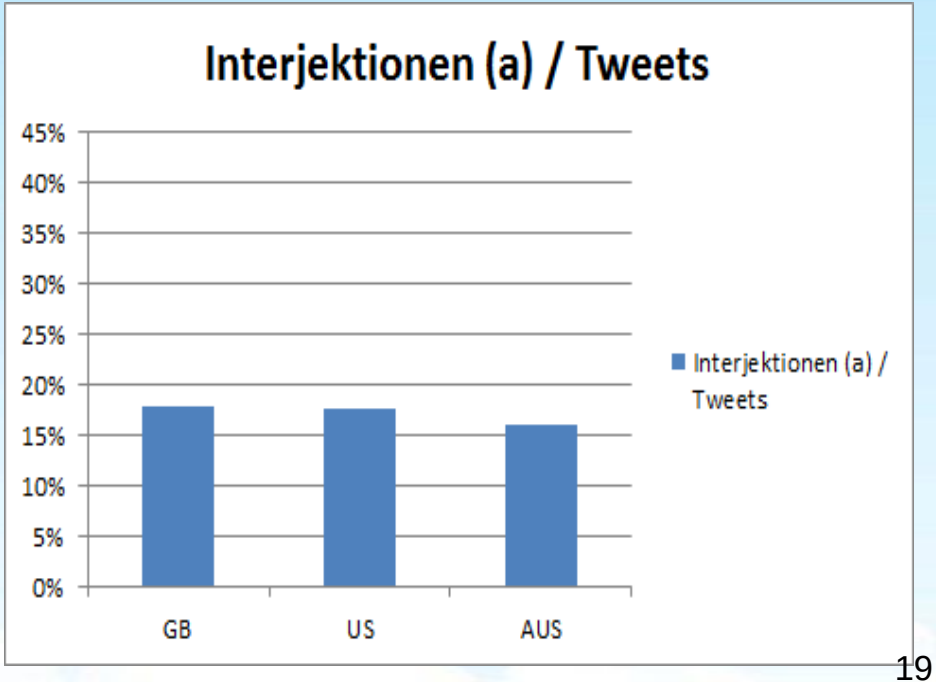
- aber Tendenz zu mehr *Sprache der Nähe* im GB-Korpus noch deutlich

Small Scale vs. Large Scale Study
Vergleich der Ergebnisse

manuelles Tagging (m)
kleines Sample



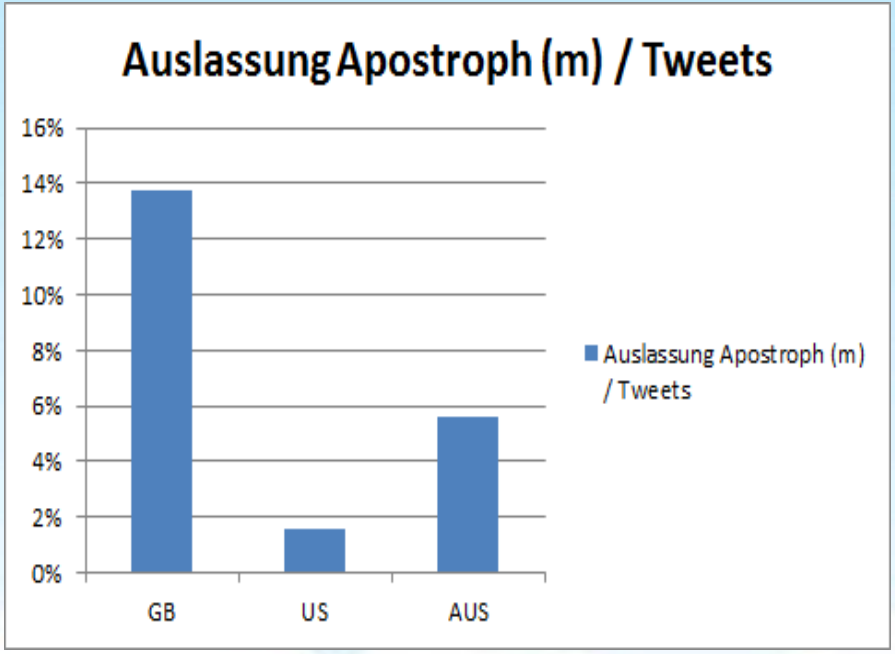
automatisches Tagging (a)
großes Sample

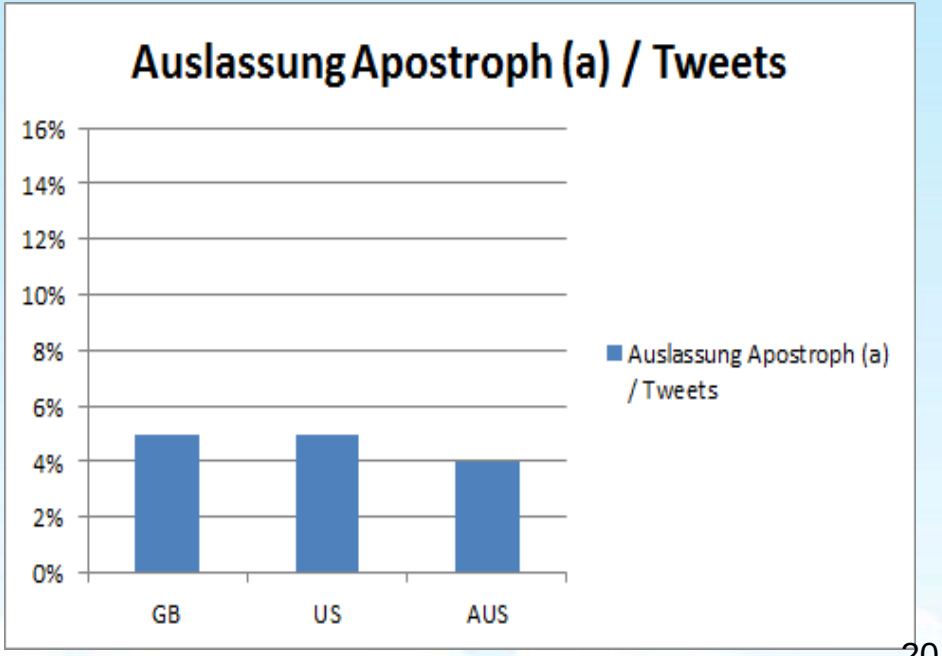


Small Scale vs. Large Scale Study
Vergleich der Ergebnisse

manuelles Tagging (m)
kleines Sample

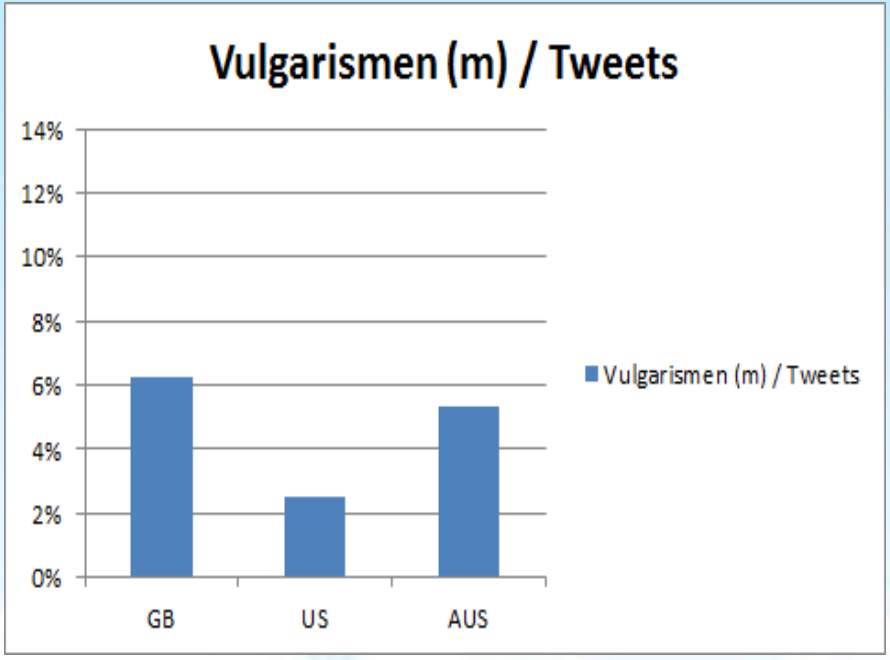
automatisches Tagging (a)
großes Sample



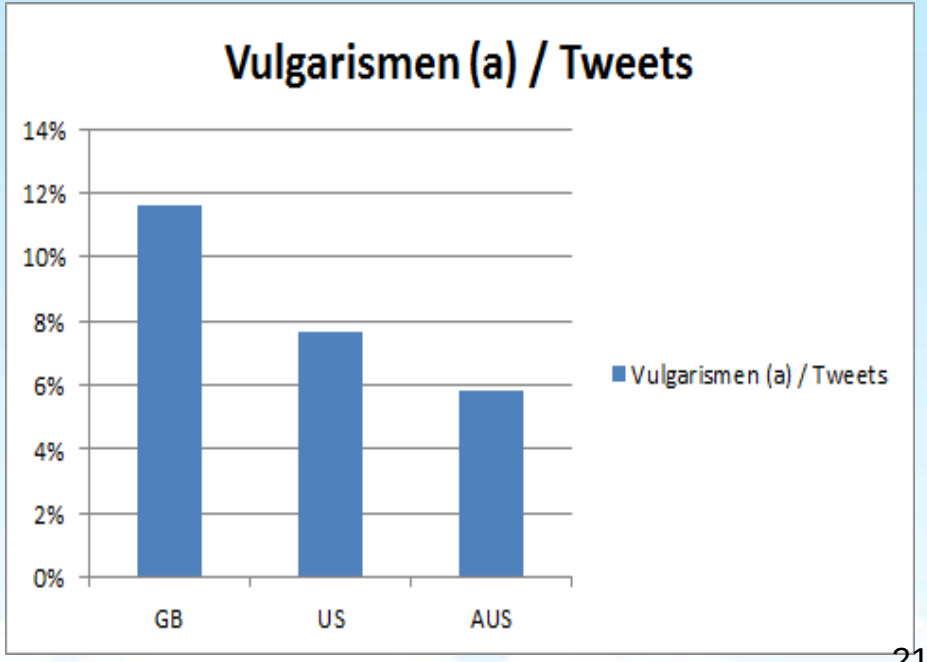


Small Scale vs. Large Scale Study
Vergleich der Ergebnisse

manuelles Tagging (m)
kleines Sample



automatisches Tagging (a)
großes Sample



Small Scale vs. Large Scale Study
Vergleichbarkeit der Ergebnisse?

- **Technische Probleme**

Sauberkeit der Korpora, Genauigkeit des Taggers (Problem der Übergeneralisierung), Genauigkeit der Auszählung

- **Diachrones Problem**

Veränderung hin zu einer Standardisierung im Substandard, technische Innovationen beeinflussen Entwicklungsprozess zusätzlich, z. B. Eingabemodus (Hardware, Autokorrektur, Emoji-Keyboard)

Small Scale vs. Large Scale Study
Vergleichbarkeit der Ergebnisse?

- **Problem der funktionalen Heterogenität**

Twitter wird funktional unterschiedlich genutzt, kann nicht EINEM Genre zugeschrieben werden, sondern verschiedenen (vgl. Twitter-Startseite)

- **Soziologisches Problem**

Die diatope Klassifizierung der Tweets nach Herkunftsland ist ein Modell, das der Realität nur bedingt gerecht wird.

besser: Analyse der CoPs als Netzwerk- oder Cluster-Analyse (vgl. Bamman, Eisenstein, Schnoebelen 2012)

References

- Bamman, David; Eisenstein, Jacob & Schnoebelen, Tyler (2012). Gender in Twitter: Styles, Stances, and Social Networks*. arXiv preprint* arXiv:1210.4567.
- Crystal, David (2011). *Internet Linguistics. A Student Guide.* Routledge*.*
- Heil, Bill & Piskorski, Mikolaj Jan (2009). Men follow Men and Nobody tweets. E-Journal: HarvardBusiness.org
- Kachru, B. (1985). Standards, codification and sociolinguistic realism: The English language in the outer circle. In: R. Quirk & H.G. Widdowson (Hrsg.) English in the World: Teaching and Learning the Language and Literatures. Cambridge University Press.
- Koch, Peter & Oesterreicher, Wulf. (1994). Funktionale Aspekte der Schriftkultur*.* In: Hartmut Günther & Otto Ludwig (Hrsg.), *Schrift und Schriftlichkeit. Ein interdisziplinäres Handbuch internationaler Forschung* (S. 587-604). Berlin/New York: De Gruyter.
- Mair, Christian (2013).The World System of Englishes: Accounting for the transnational importance of mobile and mediated vernaculars. *English World-Wide*, 34, 253-278.

Large Scale Study
Methods

general findings:
for example average length of tweet per variety
average number of word forms per variety
any other general information you think would be useful

Blogosphäre

Anteil der Twitternutzer

	2010[1]	2013[2]
AUS	--	4,09%
GB	7,20%	17,09%
USA	50,88%	50,99%

([1]webanalyticsworld.net and [2]beevolve.com)