RUB

RUHR-UNIVERSITÄT BOCHUM

SUSTAINABLE SOCIAL MEDIA RESEARCH: THE CASE OF LINGUISTICS

Tatjana Scheffler

March 2024, “After Twitter” conference


Please use the purpose made paved path provided
Speakers
Prescriptivists
Diolch
Thank you
ASDA.com

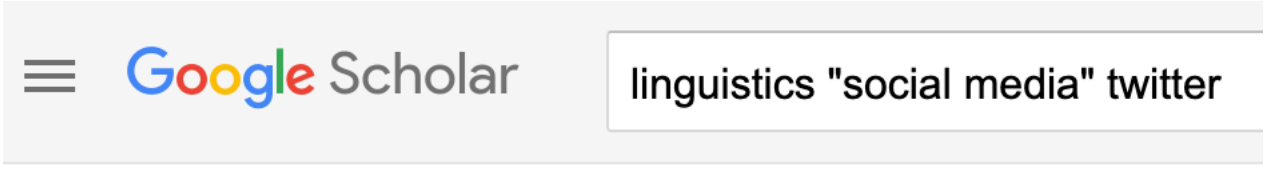
Need for social media corpora

- Classical corpora
  - News text
  - Restricted range of languages
  - 1990s
- Social media corpora
  - Very large!
  - Current topics
  - Interactive language
  - Diverse authors
  - Metadata

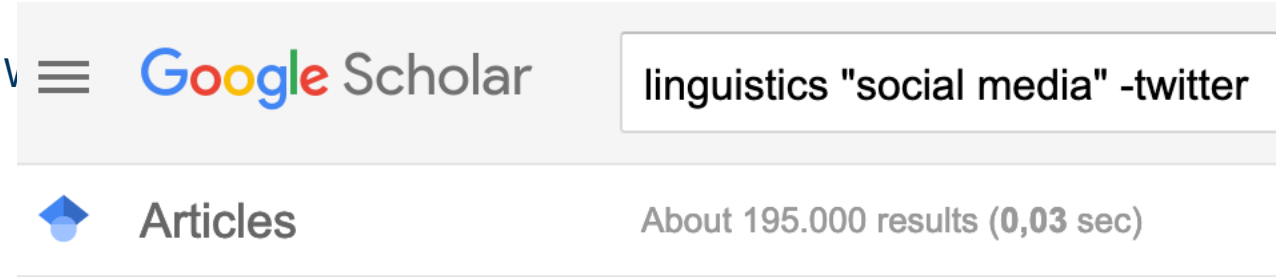
Past

Social media corpora for linguistic research

- Ad-hoc, topic-based
- Availability as main criterion
- Platforms:



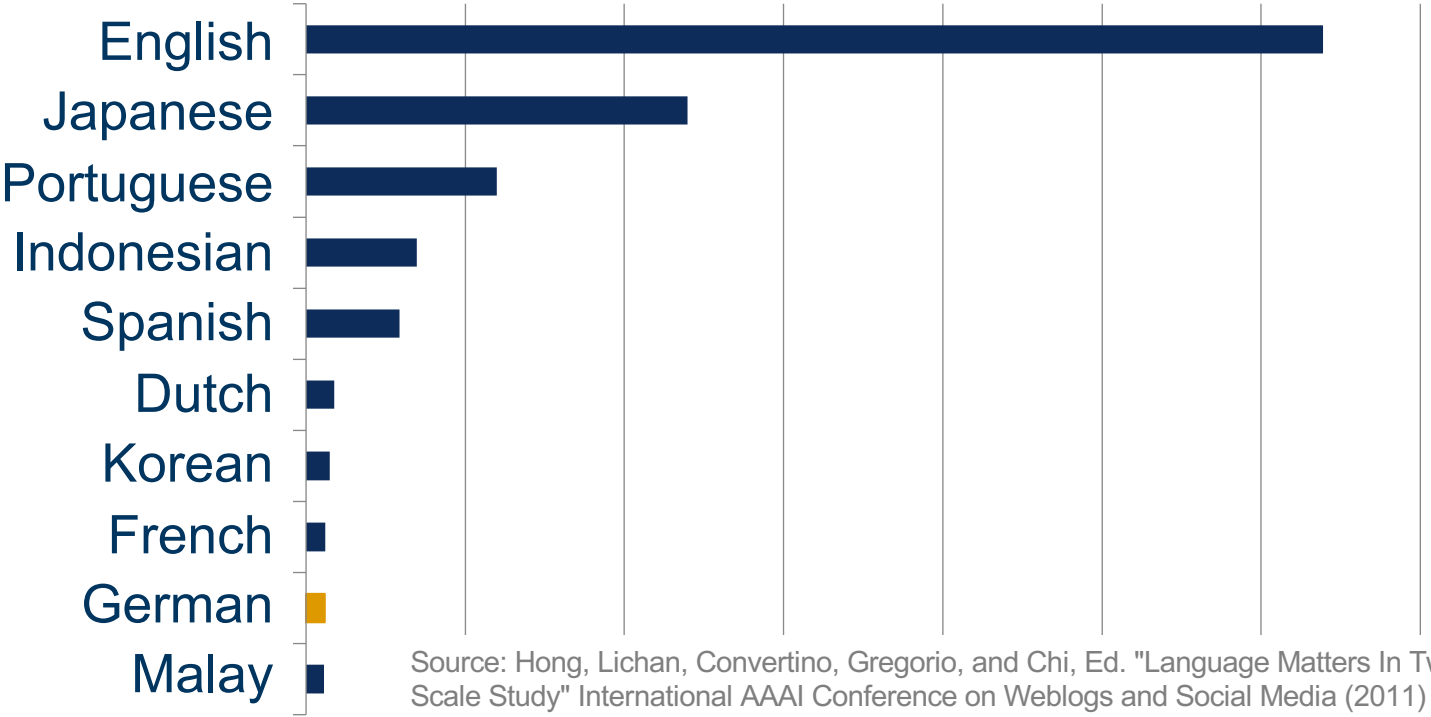
  - Comment forums (eg V
  - Blogs
  - …



Twitter-APIs for creating corpora

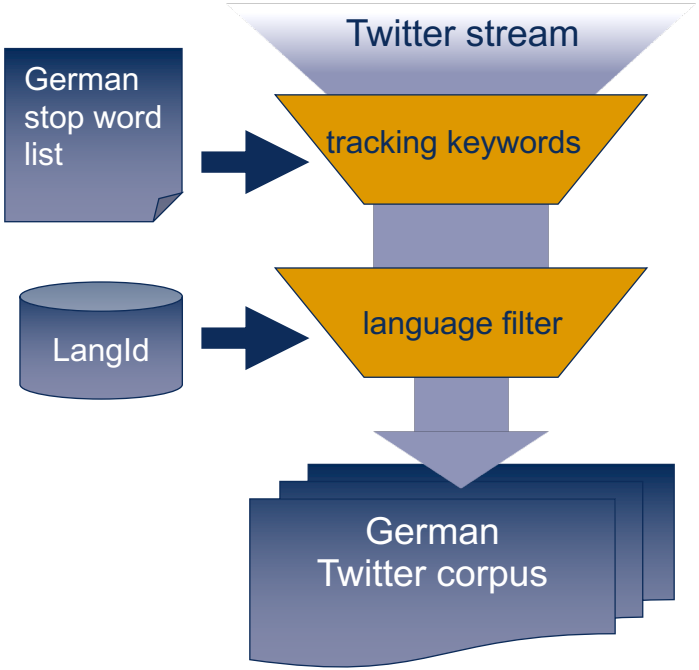
- Search API or Streaming API
- Search API: key words, up to 7 days into the past
- Streaming API:
  - real time stream of posted tweets
  - rate limitation
  - mixed (language/topic) tweets
  - filter by:
    - geo-location (location)
    - up to 5000 user ids (follow)
    - up to 400 keywords (track)

Languages on Twitter



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

Corpus creation using the streaming API

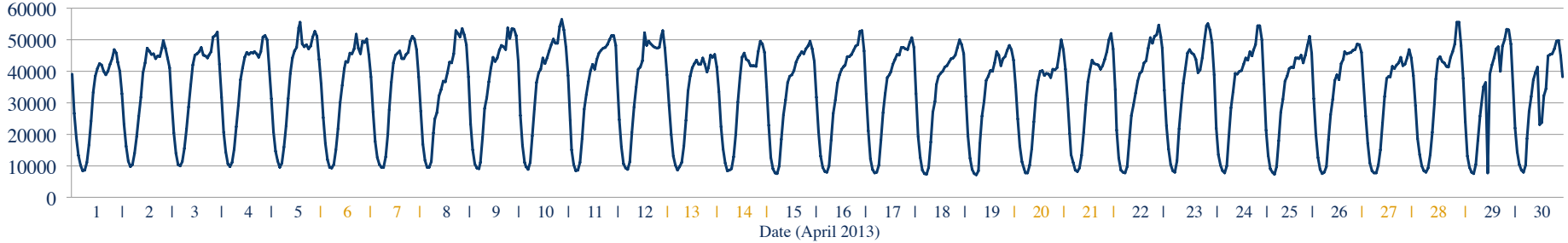


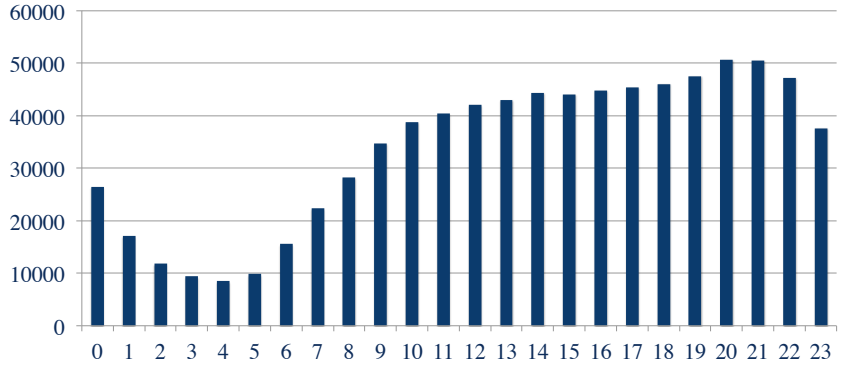
~ 500.000.000 tweets / day

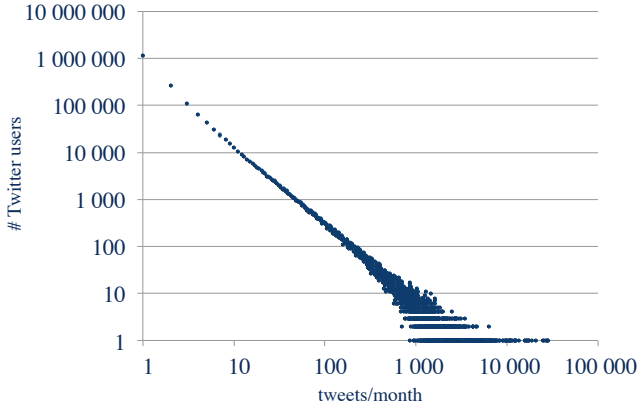
~ xx.000.000 tweets / day

~ 1.000.000 tweets / day

German Twitter data







RUHR UNIVERSITÄT BOCHUM RUB

Dealing with Twitter corpora in research

- Twitter ToS has always prohibited sharing of aggregated tweets (=corpora)!
- corpus sharing only via tweet IDs; time-consuming recrawling of individual tweets, e.g. via https://github.com/lintool/twitter-tools
- deletion of tweets and/or accounts:
  - 21.2 % of the Tweets2011 corpus were unretrievable after 9 months
- central value of (manual) annotations
- How to anonymize tweets in scientific papers?
  - removal of @handles
  - often still googleable

XXL German Twitter corpus

- XXL: eXXtra-Large German Twitter corpus
- method: (Scheffler 2014); 24 million German tweets from April 2013
- no topic/keyword based filtering
- mid 2014 – March 2023
- ~ 2 billion German tweets in total

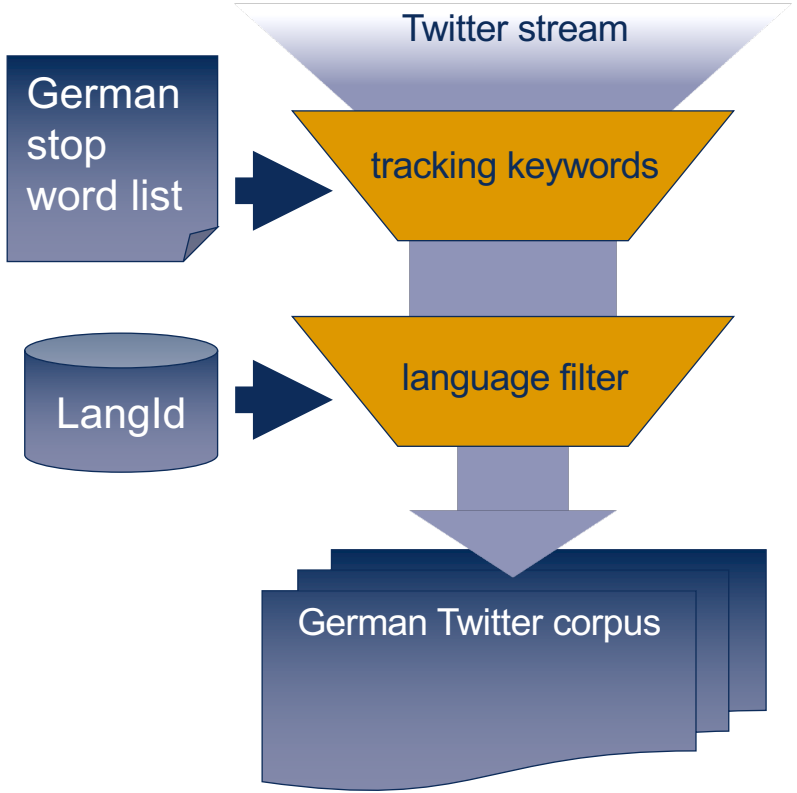
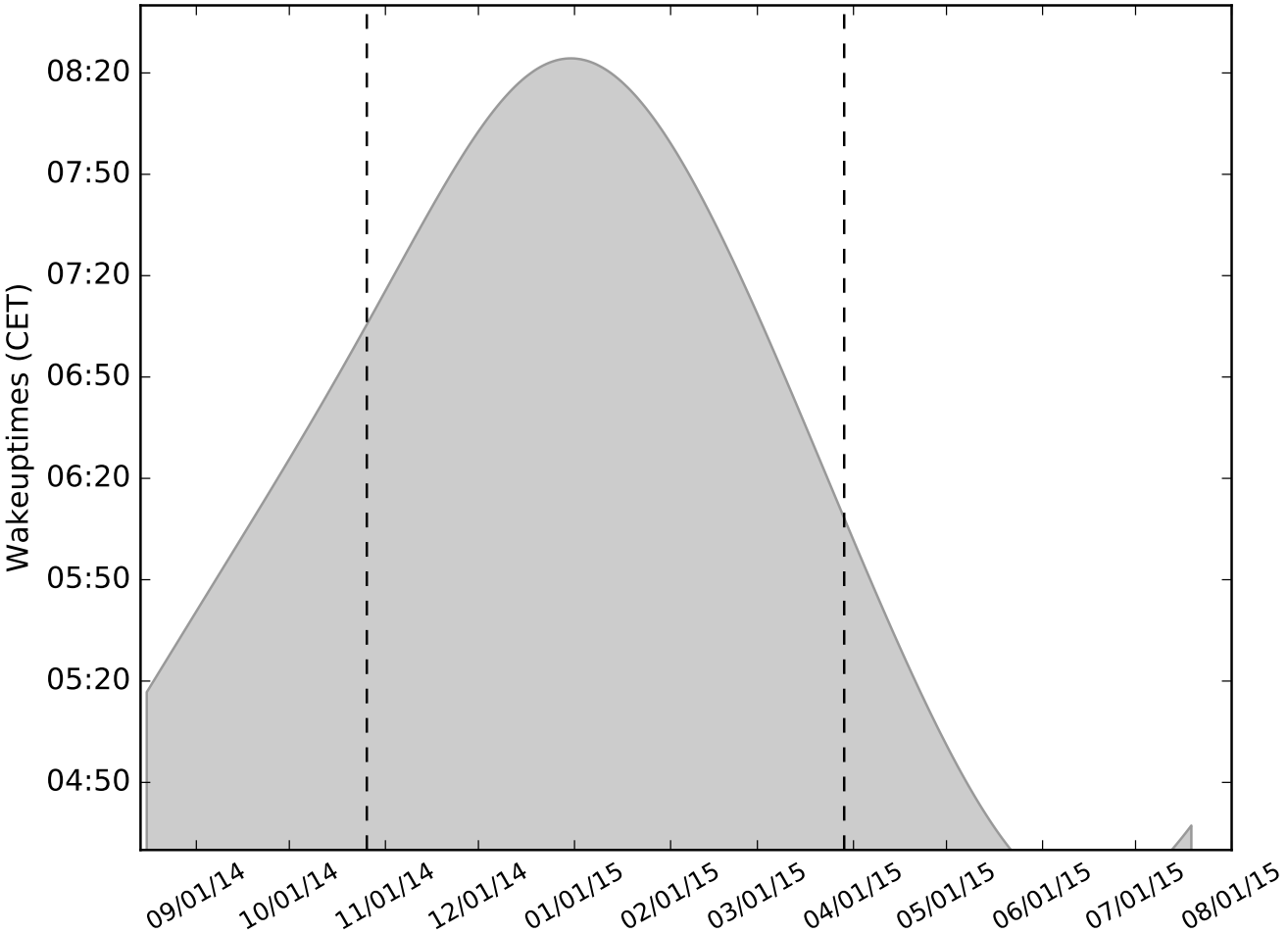
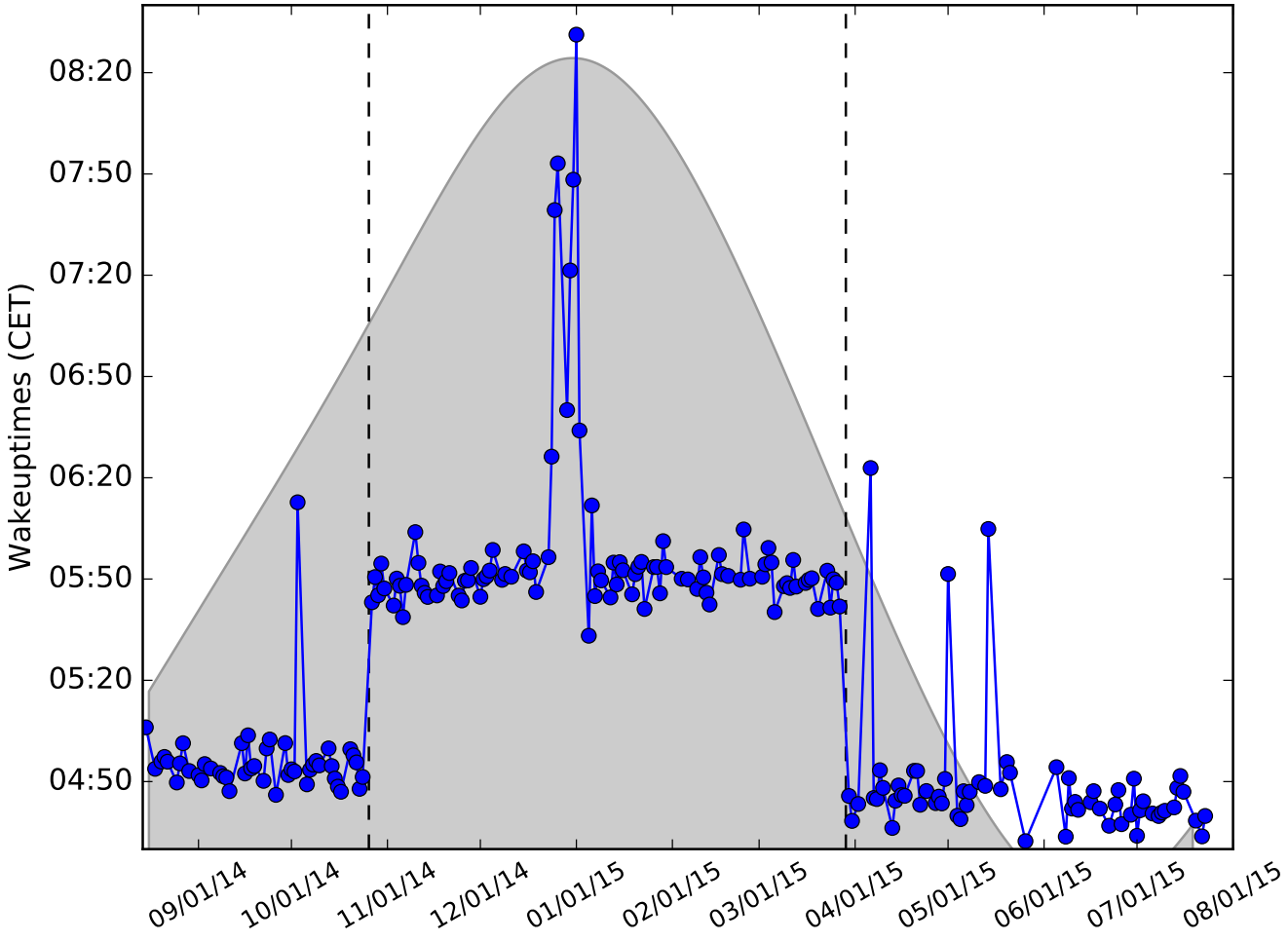
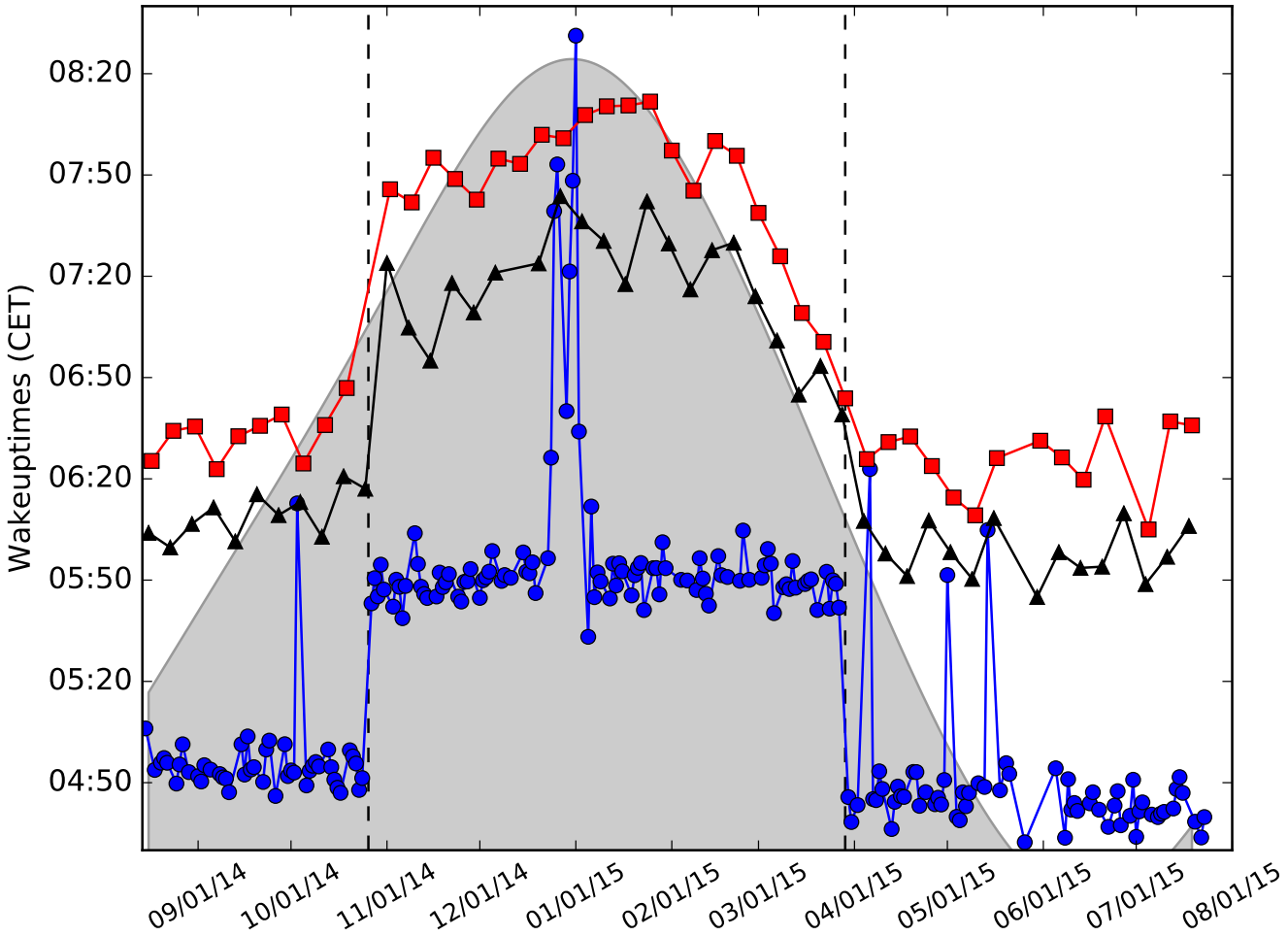


Image: Fotolia


Wakeuptimes (CET)
08:20
07:50
07:20
06:50
06:20
05:50
05:20
04:50
09/01/14
10/01/14
11/01/14
12/01/14
01/01/15
02/01/15
03/01/15
04/01/15
05/01/15
06/01/15
07/01/15
08/01/15


Wakeuptimes (CET)
08:20
07:50
07:20
06:50
06:20
05:50
05:20
04:50
09/01/14
10/01/14
11/01/14
12/01/14
01/01/15
02/01/15
03/01/15
04/01/15
05/01/15
06/01/15
07/01/15
08/01/15


Wakeuptimes (CET)
08:20
07:50
07:20
06:50
06:20
05:50
05:20
04:50
09/01/14
10/01/14
11/01/14
12/01/14
01/01/15
02/01/15
03/01/15
04/01/15
05/01/15
06/01/15
07/01/15
08/01/15

Present

Bochum CMC Corpora

- German Twitter snapshot (April, 2013)
- XXL: eXXtra-Large German Twitter corpus
  - 2014–2023
  - majority of German tweets
  - no topic/keyword based filtering
  - ~ 2 billion tweets

- ChrisTof: religious forum data
- PARADISE: parallel blog posts & podcasts

- TwiBloCoP: Twitter/blogs from the same authors

TwiBloCoP – Privacy

- Collection of data follows §60d UrhG
- 2021: Inform authors, give option to withdraw consent (Opt-Out)
- 62 authors, 50 could be contacted:

  3 refusals, 41 agree, 6 agree after answering questions

Manual Pseudonymization

- personal name, blog name, email, place, @username, url, phone number

„Plowed sidewalks!
In Frankfurt they flatten the
snow cover and then they spread
gravel (or whatever it's called) over it
like crazy.“

„Plowed sidewalks!
In [PLACE] they flatten the
snow cover and then they spread
gravel (or whatever it's called) over it
like crazy.“

CMC Linguistics

The medium is not the message

Individual level register variation in blogs vs. tweets

DE GRUYTER MOUTON

Luca Bevacqua* and Tatjana Scheffler

Form variation of pronoı written English

A corpus study in Twitter and iWeb

The Telegram Chronicles of Online Harm

Journal of Open Humanities Data

n[2] & Hannah Seemann[1]
Jniversität Marburg

along many axes, including ffordances, and others. In this st be distinguished within social l as an important factor indepen- able linguistic phenomena. We

Language Sciences 96 (2023) 101535

Contents lists available at ScienceDirect

Language Sciences

journal homepage: www.elsevier.com/locate/langsci

Tracing and classifying German intensifiers via information theory

Tatjana Scheffler[a,*], Michael Richter[b], Roeland van Hout[c]

[a] Ruhr-Universität Bochum, Fakultät für Philologie, Germanistisches Institut, Universitätsstraße 150, 44780 Bochum, Germany
[b] Universität Leipzig, Germany
[c] Radboud University, Netherlands

E GRUYTER MOUTON Corpus Linguistics and Ling. Theory 2022; 18(1): 1–31

ulia Clausen* and Tatjana Scheffler

corpus-based analysis of meaning ariations in German tag questions

vidence from spoken and written conversational orpora

tps://doi.org/10.1515/cllt-2019-0060

bstract: This paper addresses semantic/pragmatic variability of tag questions German and makes three main contributions. First, we document the pre- alence and variety of question tags in German across three different types of onversational corpora. Second, by annotating question tags according to their ntactic and semantic context, discourse function, and pragmatic effect, we

Future

Desiderata

- Researchers:
  - long-term access
  - option to access and add/edit annotations
  - context, metadata
  - size
- Authors:
  - protection of personal data
  - informed consent
  - benefit from research results
- Platforms:
  - monetization

Strategies for collecting social media corpora

- data donation
  - scale: 281 million tweets (2022 subsection of XXL) compared to 38,000 WhatsApp posts (MoCoDa2)
  - age
  - manual processing
- opportunistic archiving
  - web-based
  - independent of topic and specific incidents
  - provide derived formats
  - develop sharing mechanisms for raw & annotated data
- effective pseudonymization

THANK YOU!

Tatjana Scheffler

tatjana.scheffler@rub.de

and the Digital Forensic Linguistics team

SFB 1287 Limits of Variability in Language Cognitive, Grammatical, and Social Aspects

ViCom

SFB Mo Re Metaphors of Religion

Faktenforum

VIRTUELLE Lebenswelten

RUHR UNIVERSITÄT BOCHUM

RUB

Image references:

- Title: Lewis Ogden, https://www.flickr.com/photos/bitsfrombytes/43617178595
- text by Alice Design from Noun Project (CC BY 3.0)
- Scroll by Designing Hub from Noun Project (CC BY 3.0)
- Future by sumarni from Noun Project (CC BY 3.0)
- User, social media, shirt, tee shirt: freepik via Flaticon.com

Selected paper references:

Tatjana Scheffler, Hannah Seemann, Lesley-Ann Kern. The medium is not the message: Individual level register variation in blogs vs. tweets. Register Studies 4(2). 2022. https://doi.org/10.1075/rs.22009.sch

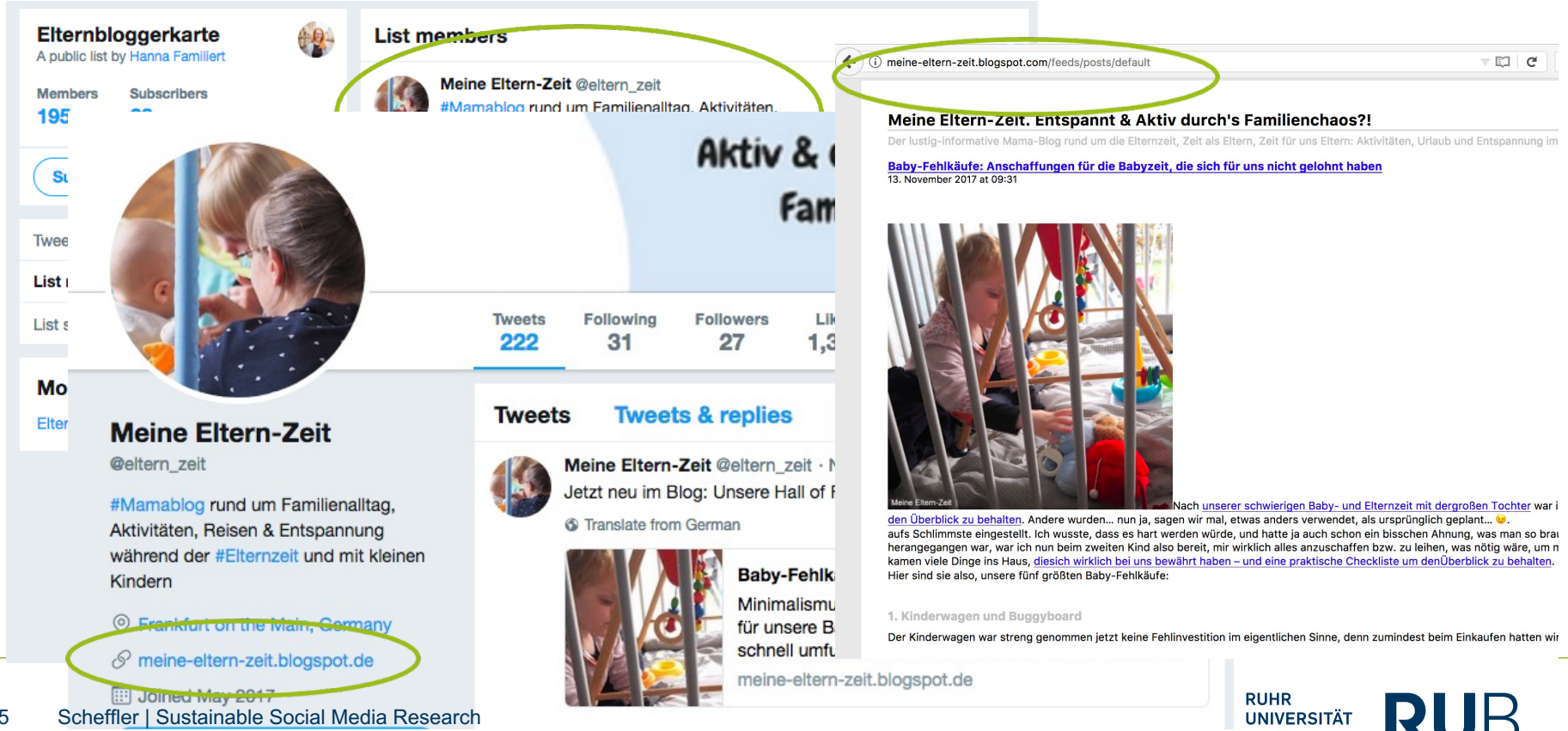
Tatjana Scheffler, Lesley-Ann Kern and Hannah Seemann. Individuelle linguistische Variabilität in sozialen Medien. In: M. Kupietz/T. Schmidt (eds.), Neue Entwicklungen in der Korpuslandschaft der Germanistik: Beiträge zur IDS-Methodenmesse 2022. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache (CLIP) 11). Tübingen: Narr. 2023.

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

Tatjana Scheffler and Christopher C.M. Kyba. Measuring Social Jetlag in Twitter Data. In: Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016), AAAI, Köln, Germany. 2016.

Schöch, Christof, Frederic Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann & Jörg Röpke. 2020. Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. Zeitschrift für digitale Geisteswissenschaften.

Creating a cross-channel corpus



RUHR UNIVERSITÄT BOCHUM RUB

TwiBloCoP

- Twitter and Blog Corpus – Parenting
- http://tiny.cc/twiblocop
- Collection period: Oct. 2016–Feb. 2017
- Topics: Family life and parenting
- Raw text, tokenized, TEI XML format

	blog posts	tweets
users	44	44
posts	468	81,440
tokens	~360,000	~1,200,000

TwiBloCoP

(1) ‘Children are our mirrors. If you want to change your child, change YOUR behavior, not the child’s. My son has these tantrums all the time. Regularly. Then it is very difficult to get him out of it. And that is exactly what I would like to do. […] Hm. At some point I asked myself why these fits upset me so much.’
[blog-4421-10]

(2) Alarm rang every 5 minutes since 6 am. Got up right before 8. Great. Worked like a charm🙈
[tweets-7291]

Linguistic annotation: register

(3) ‘Am Dienstag ist Valentinstag ! ❤️ An unserem #PinterestSonntag hat [NAME] die herzigsten DIYs für Omas und Tanten ... [URL]’

‘Tuesday is Valentine’s Day ! ❤️ For #PinterestSunday , [NAME] wrote about the sweetest DIYs for moms and aunts … [URL]’

[tweets-2191]

👉 **informative register**

(4) ‘Jungs schlafen . Gatte ist aus ... Und ihr so ? [URL]’

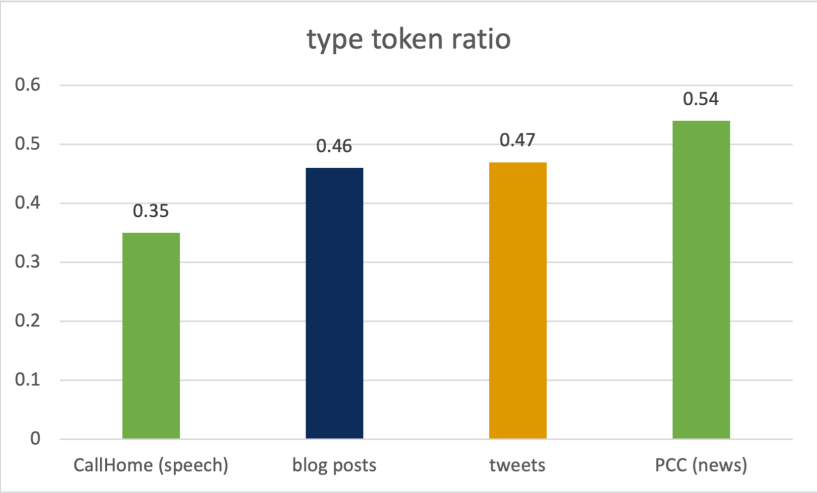
‘Boys are asleep . Husband is out … What about you all ? [URL]’

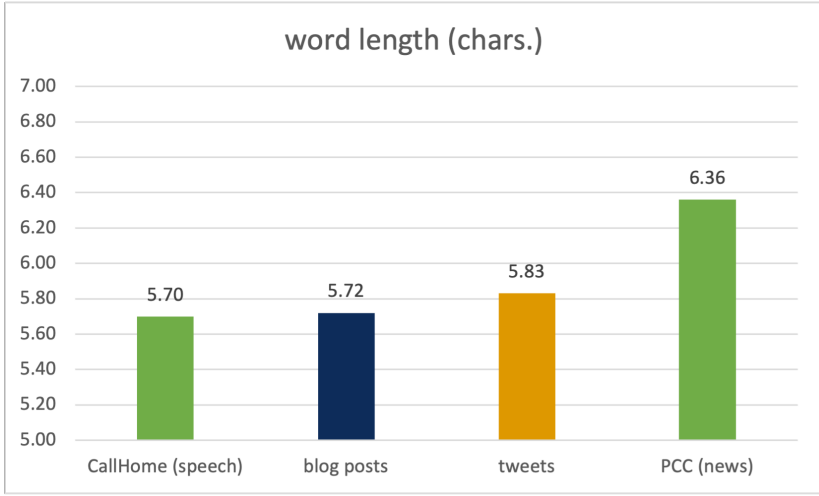
[tweets-2191]

👉 **narrative register**

Characteristics of CMC data

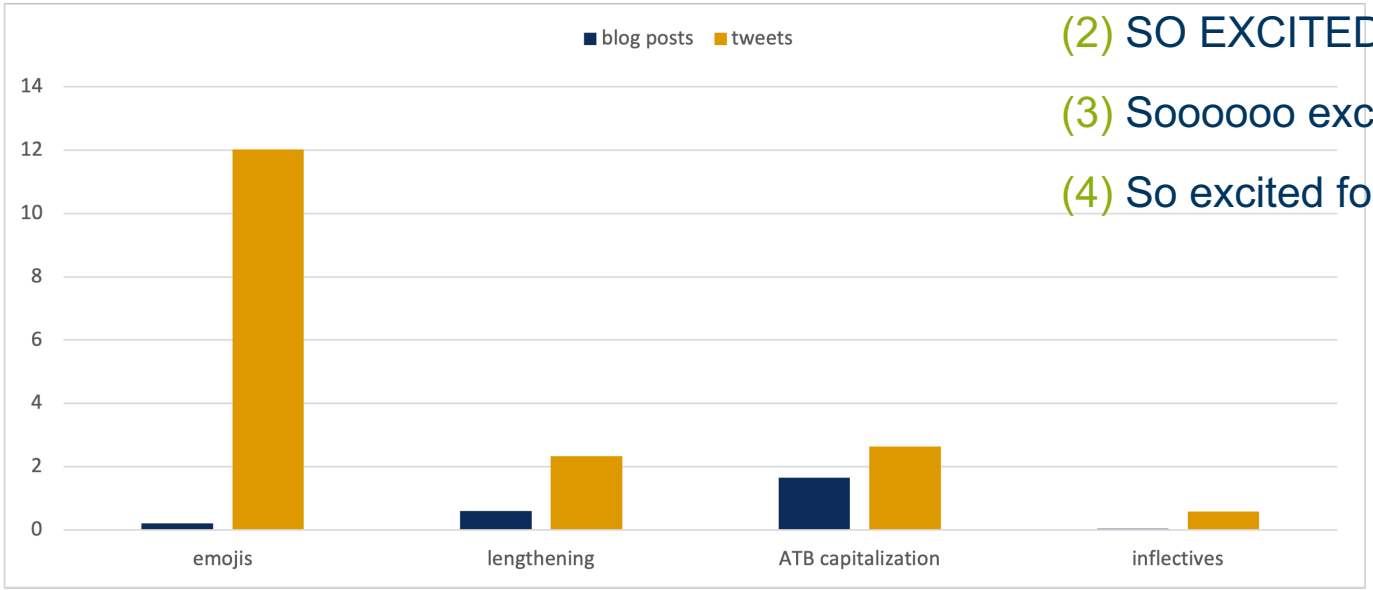
- distinct from typical speech and writing, but can share features of either
- in measures of complexity, (in)formality both blogs/Twitter differ from (news) text
- but Twitter more interactive, and contains “innovative” social-media specific features





Non-standard items

- Frequencies per 1000 tokens:



(1) So excited for CMC-Corpora 🤗

(2) SO EXCITED FOR CMC-CORPORA!

(3) Soooooo excited for CMC-Corpora!

(4) So excited for CMC-Corpora *freu*