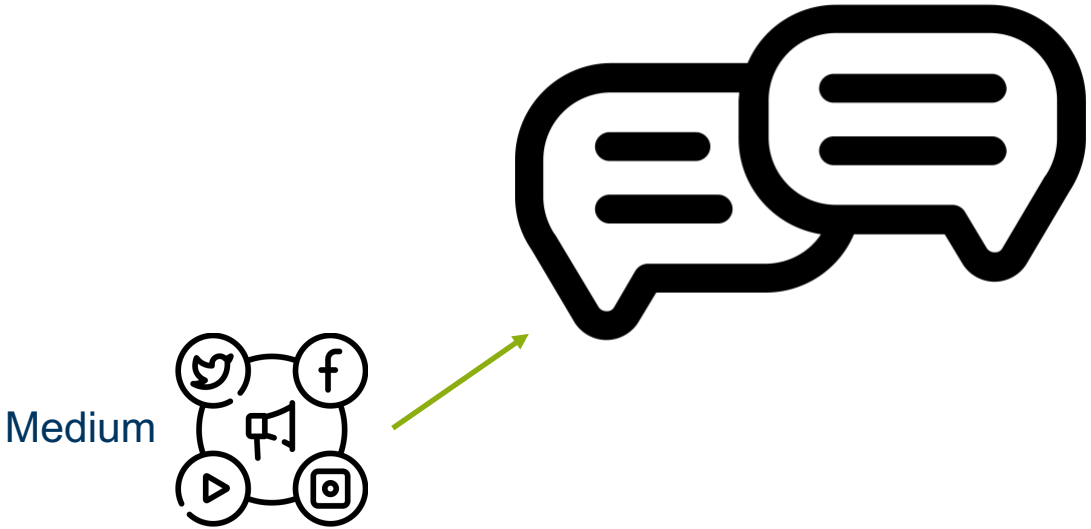
RUB

RUHR-UNIVERSITÄT BOCHUM

INDIVIDUAL LINGUISTIC VARIABILITY IN SOCIAL MEDIA

Tatjana Scheffler

September 14, 2023, CMC Corpora


Medium

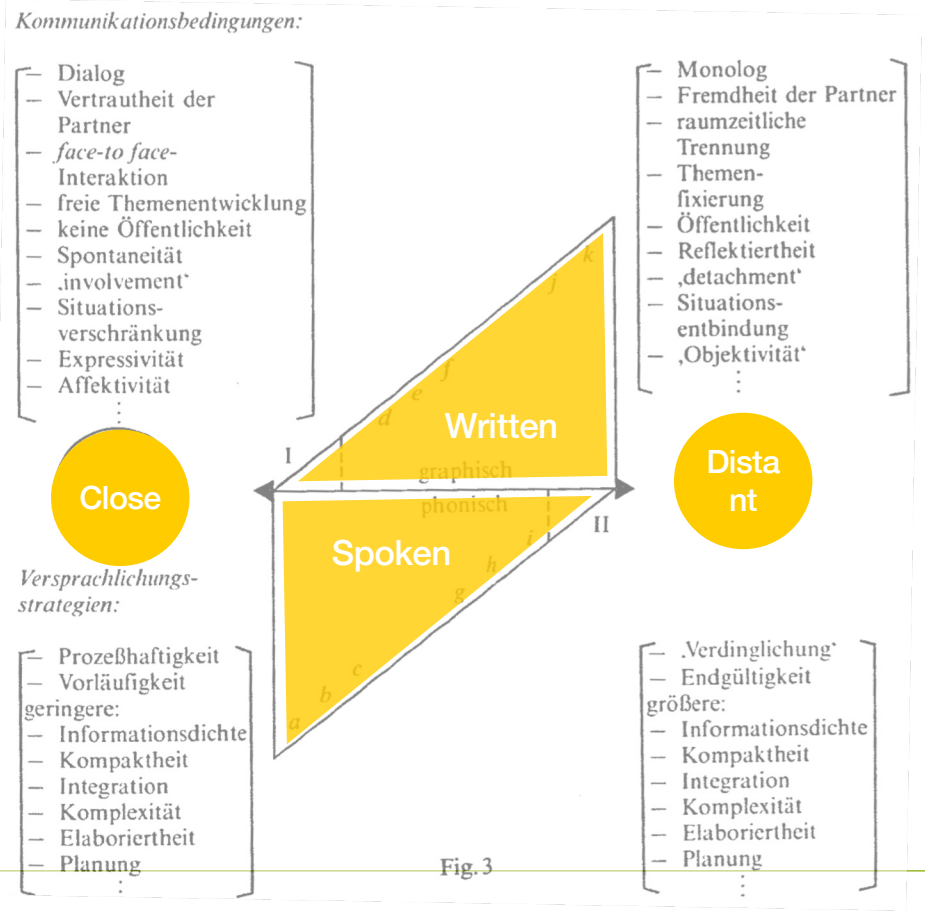


Fig. 3

Koch/Österreicher (1985)

Differences between (social) media

- Conceptual orality continuum (Koch/Oesterreicher, 1985)

- No single parameter distinguishes between “spoken” and “written” language, many features/dimensions need to be considered (Biber, 1993)
- Social media (text messages) exhibit a mix of innovative or speech-like and conservative (written-like) features (Tagliamonte/Denis, 2008)
  - non-standard features, personal pronouns
  - intensifiers
  - mix of innovative and conservative variants, often in the same turn/sentence

Non-standard items

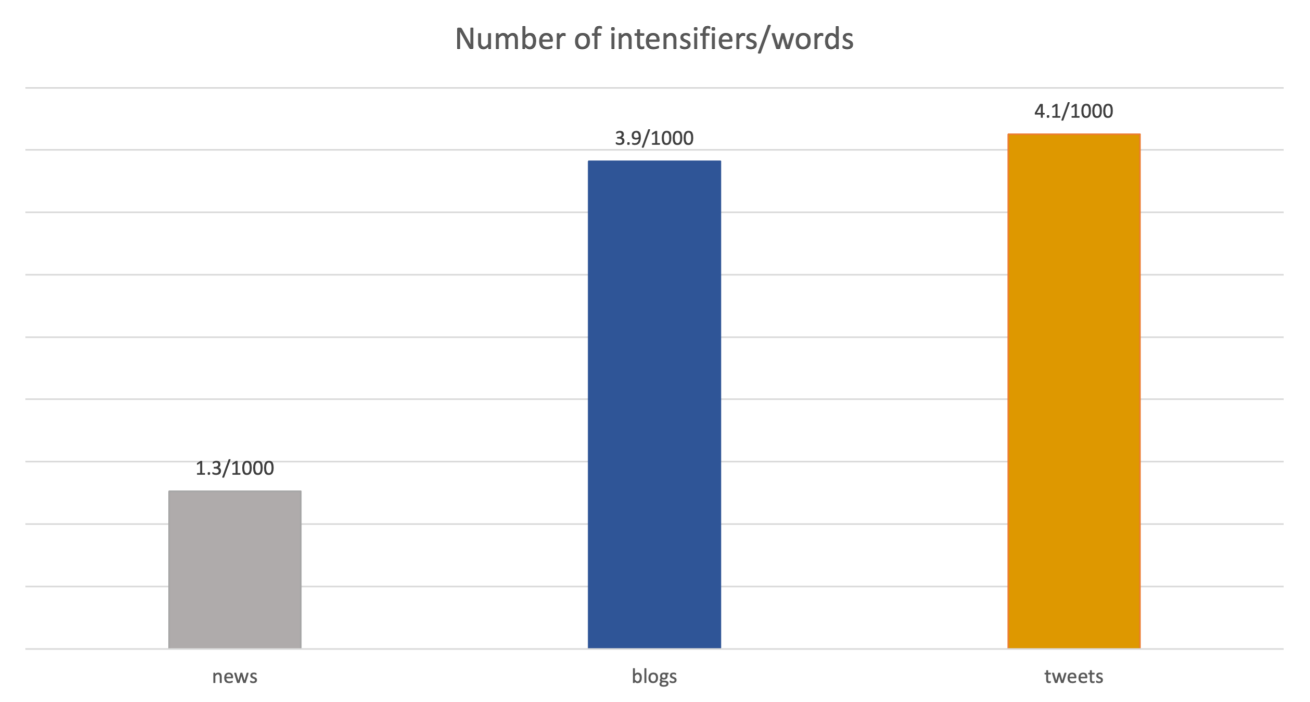
(1) So excited for CMC-Corpora 🤗

(2) SO EXCITED FOR CMC-CORPORA!

(3) Soooooo excited for CMC-Corpora!

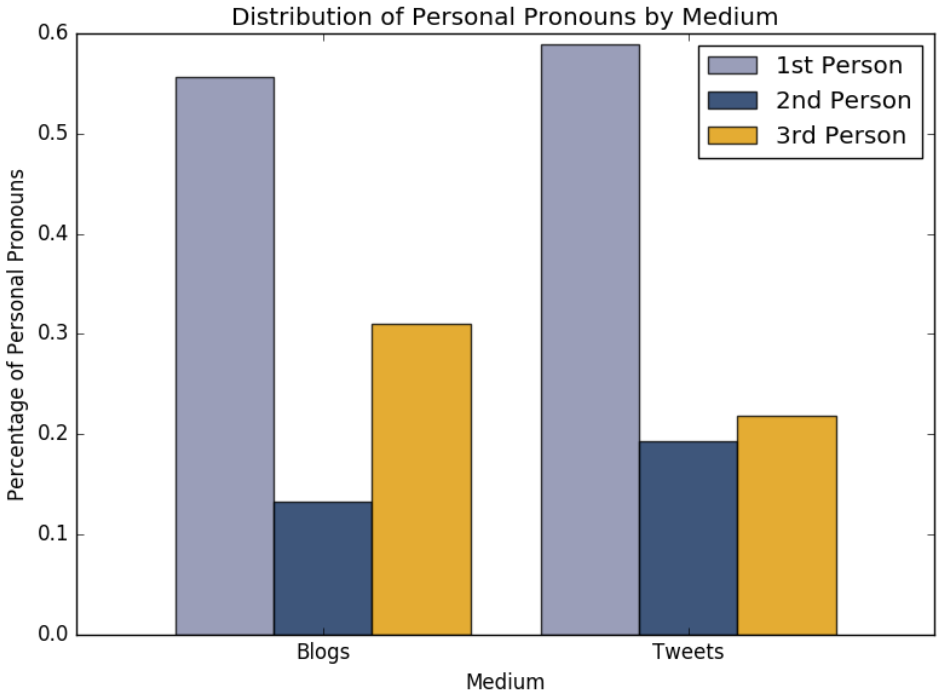
(4) So excited for CMC-Corpora *freu*

Intensifiers



Personal pronouns

- Potsdam Commentary Corpus (German news): 88% third person pronouns
- Twitter / blogs:



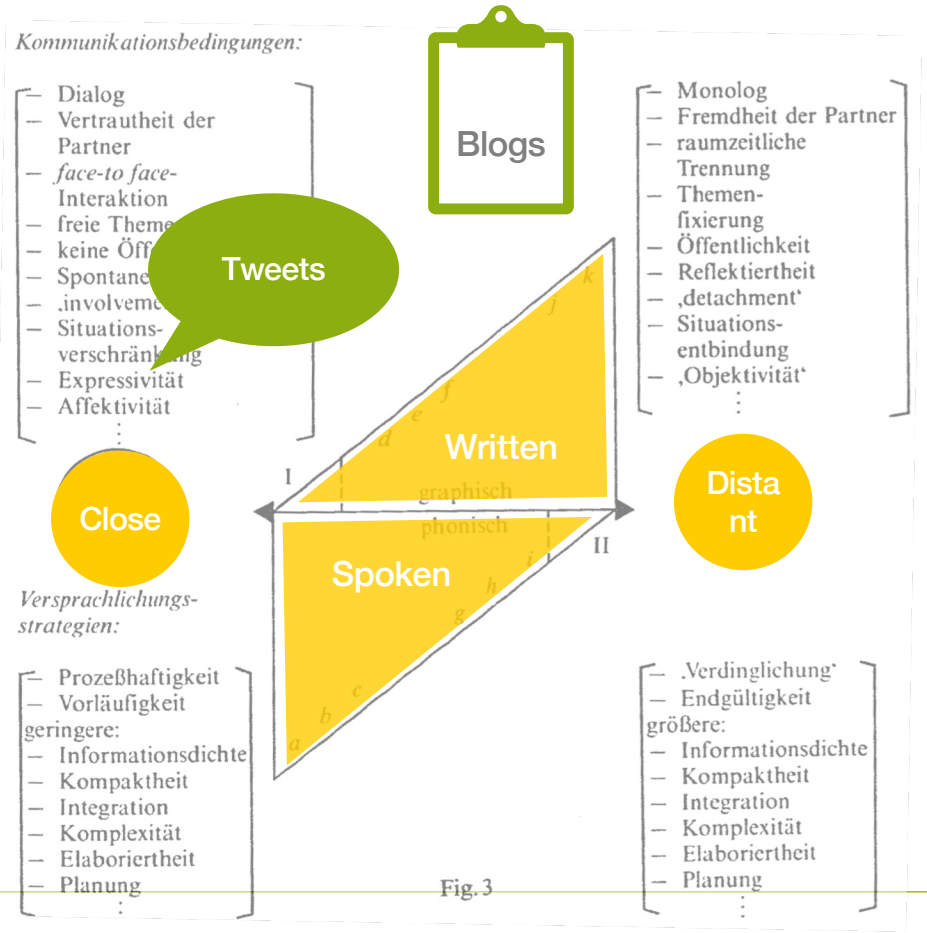
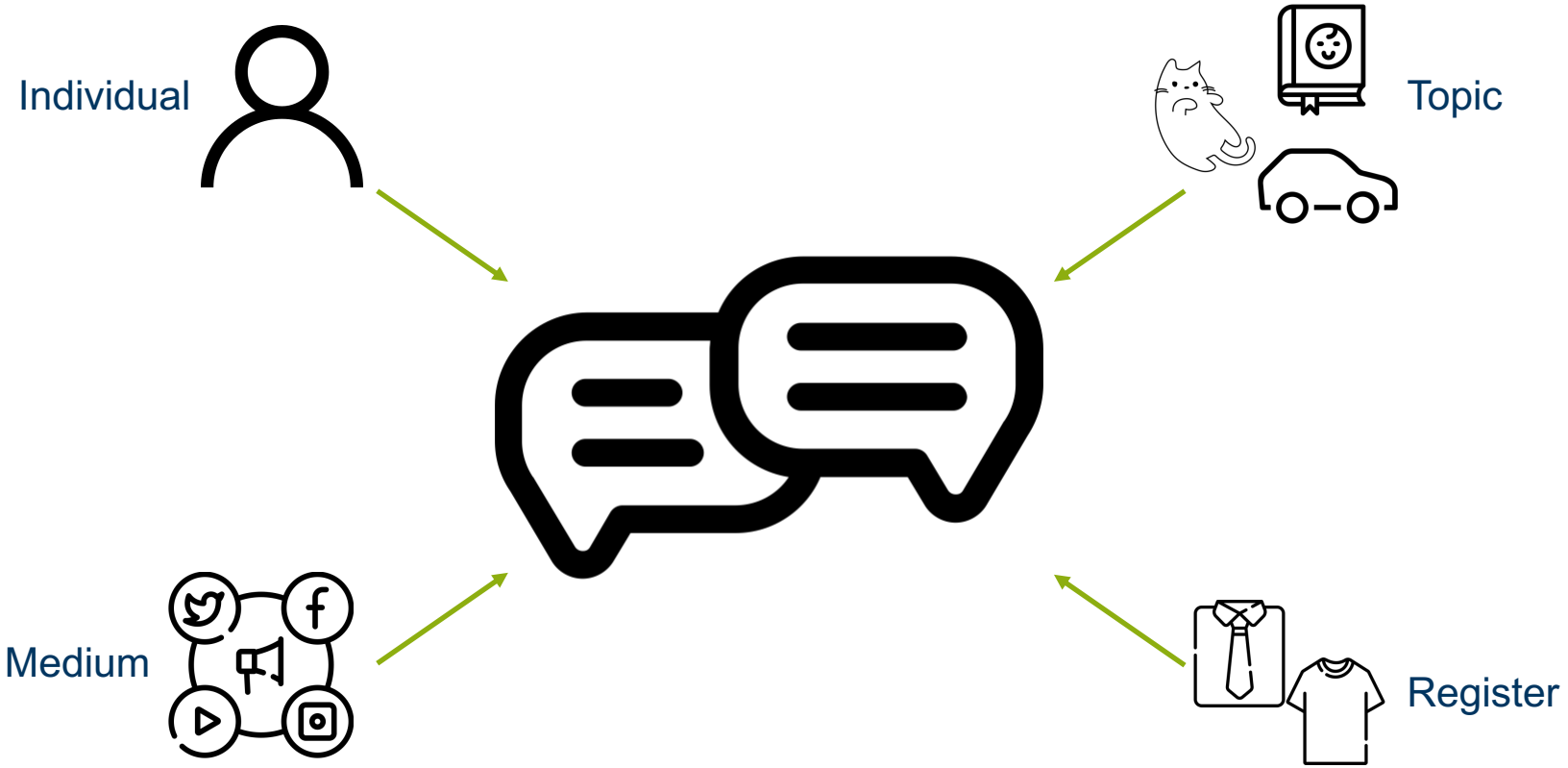


Fig. 3

Koch/Österreicher (1985)


Individual
Topic
Medium
Register

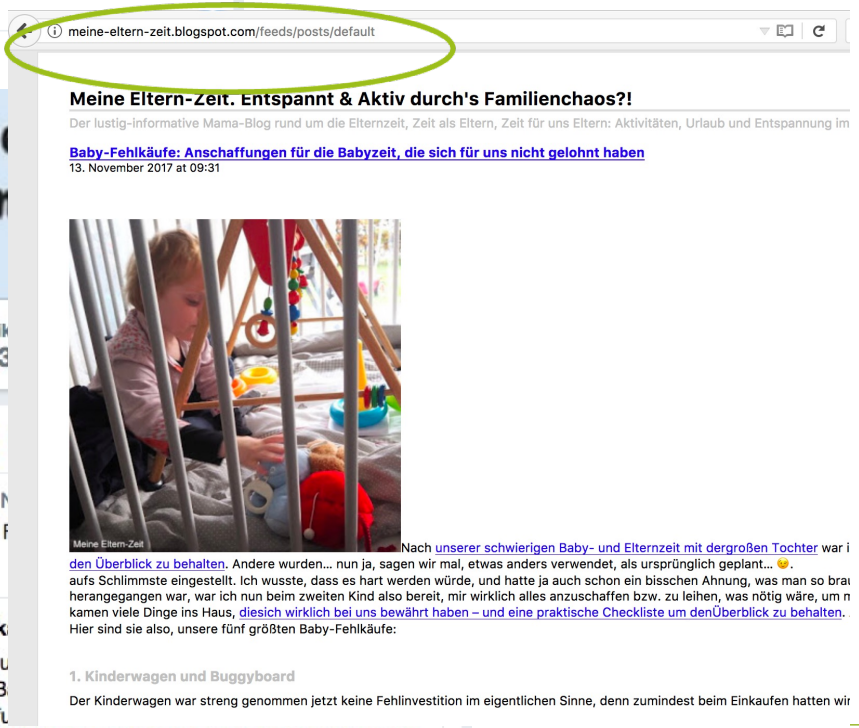
Constructing CMC Corpora

CMC Corpora

Principles:

1. Sustainable
2. Ethical
3. Open

Creating a cross-channel corpus



RUHR UNIVERSITÄT BOCHUM RUB

Privacy

- Collection of data follows §60d UrhG
- Informing 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 Mannheim 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.“

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

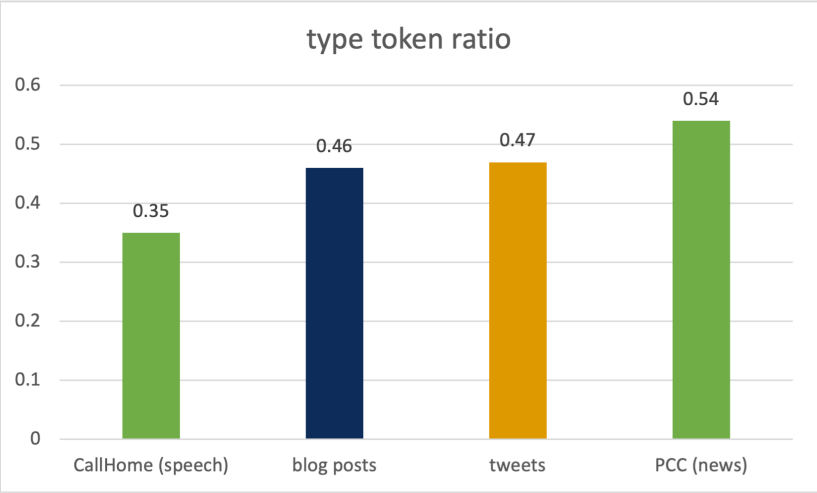
TwiBloCoP

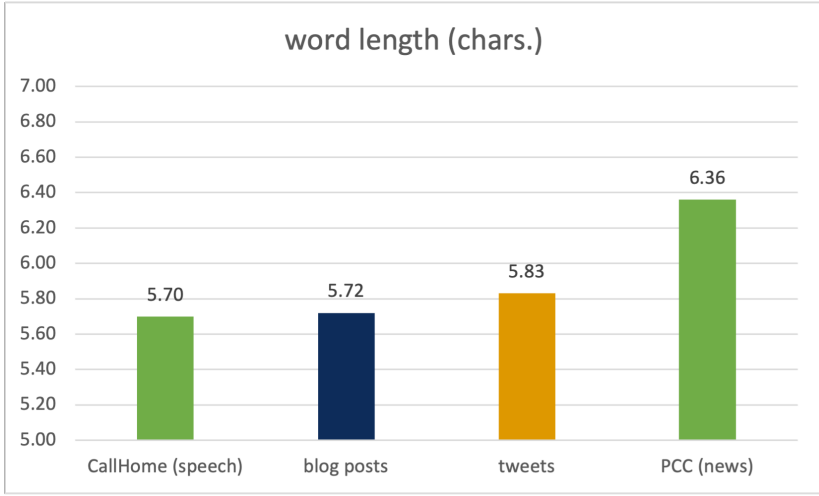
(1) ‘Children are our mirrors. If you want to change your child, change YOUR behavior, not the child’s. My son has these tantrums all the time. Regularly. Then it is very difficult to get him out of it. And that is exactly what I would like to do. […] Hm. At some point I asked myself why these fits upset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

Characteristics of CMC data

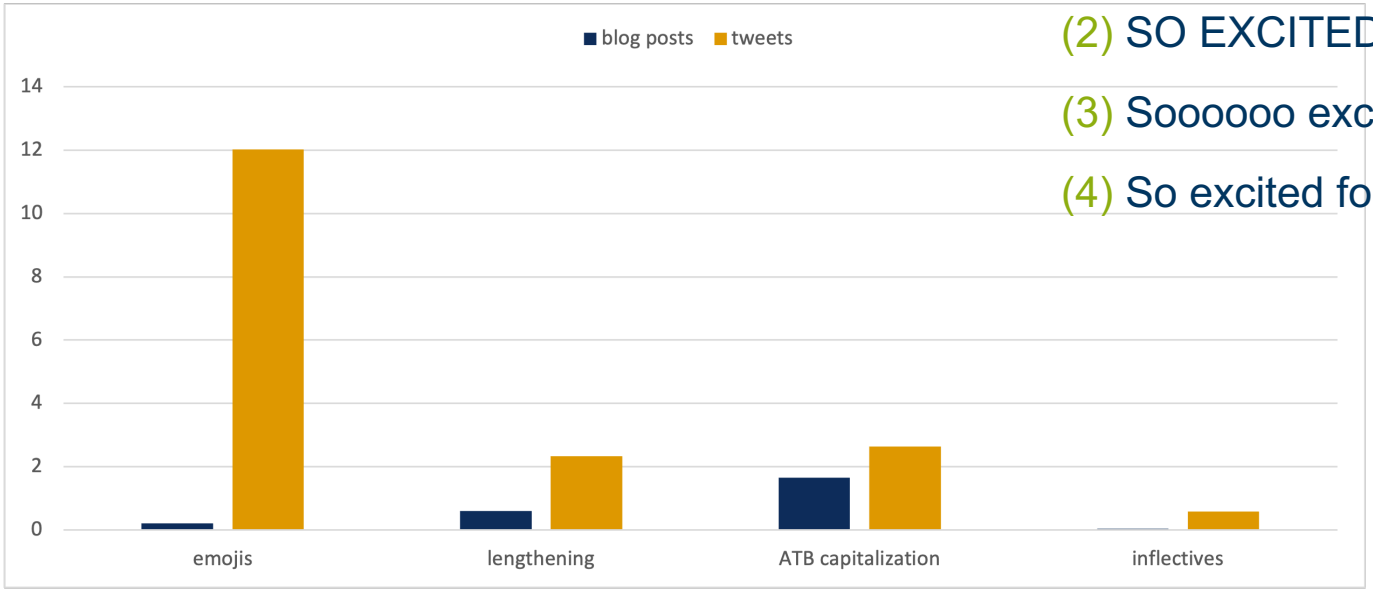
- Distinct from typical speech and writing, but can share features of either
- In measures of complexity, (in)formality both blogs/Twitter differ from (news) text
- But Twitter more interactive, and contains “innovative” social-media specific features





Non-standard items

- Frequencies per 1000 tokens:

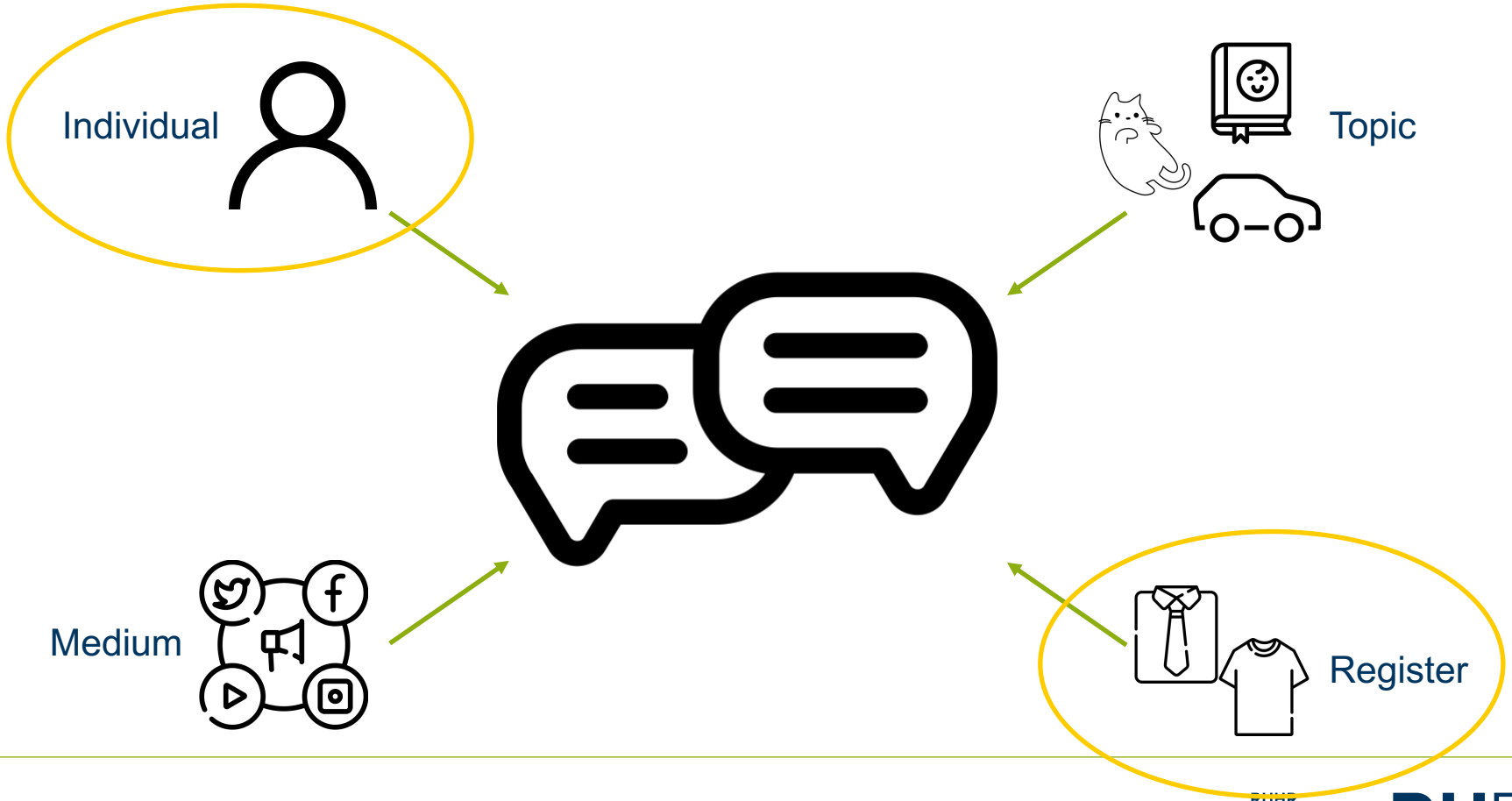


(1) So excited for CMC-Corpora 🤗

(2) SO EXCITED FOR CMC-CORPORA!

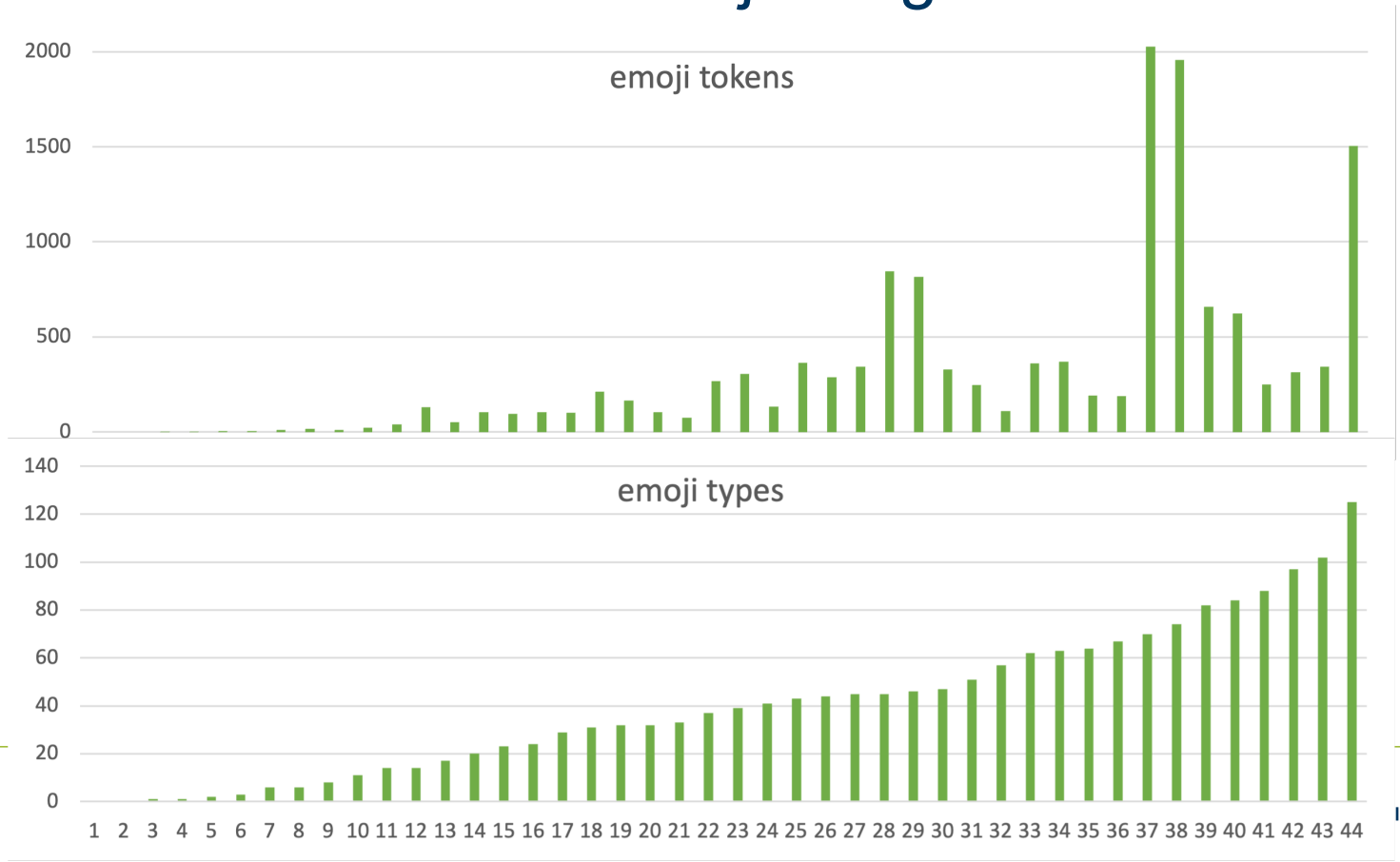
(3) Soooooo excited for CMC-Corpora!

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

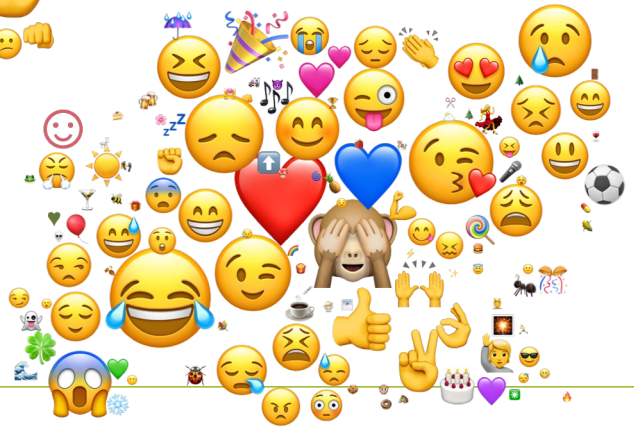

Individual
Topic
Medium
Register

Individual variation

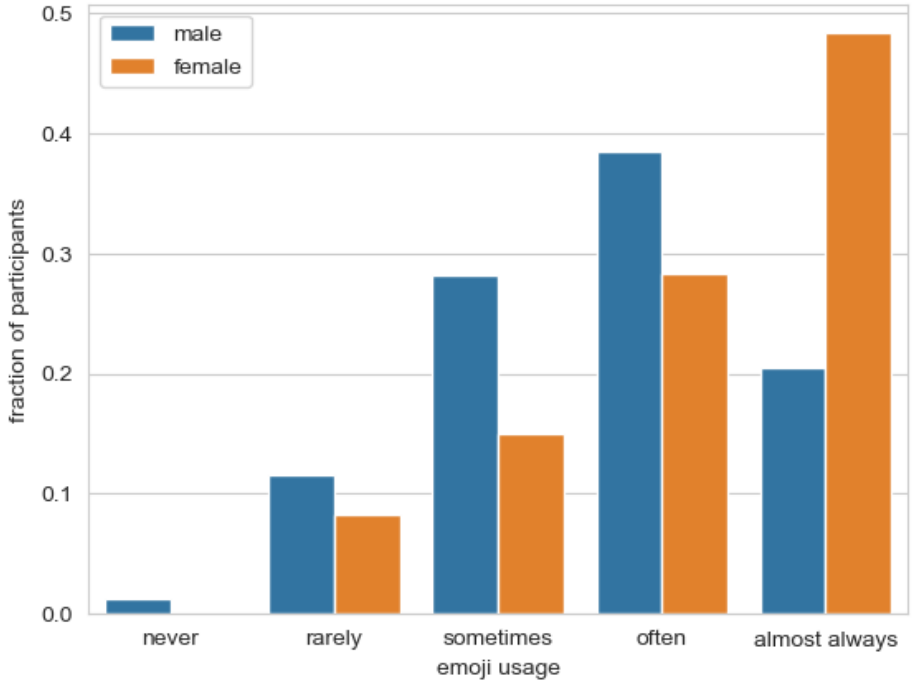
Individual variation in emoji usage



Individual variation in emoji usage

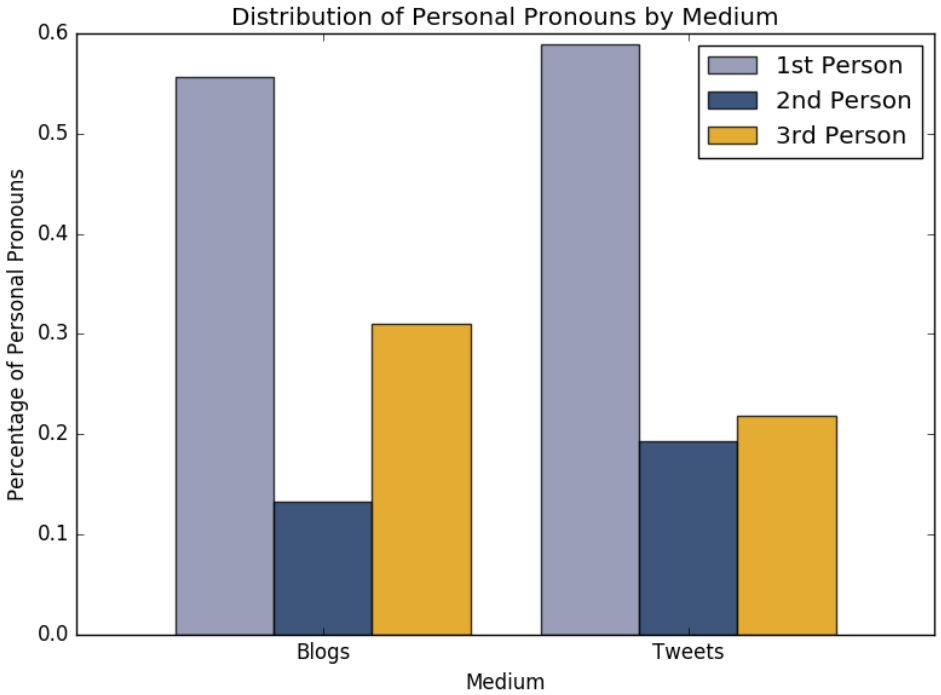


Emoji usage by gender

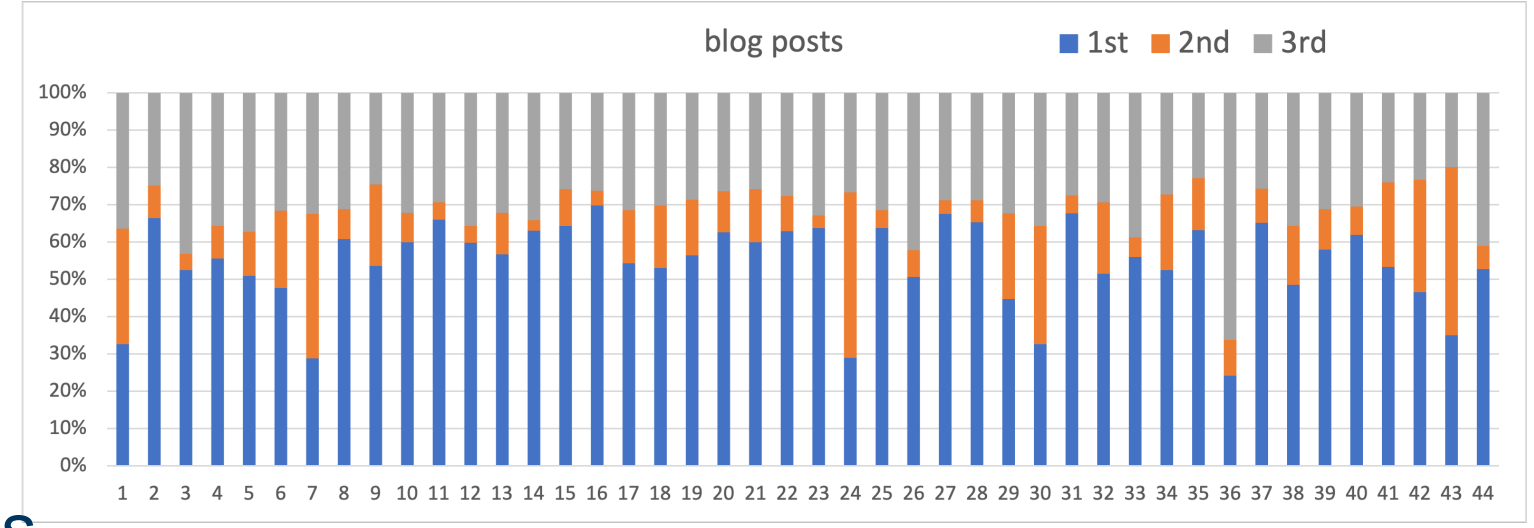


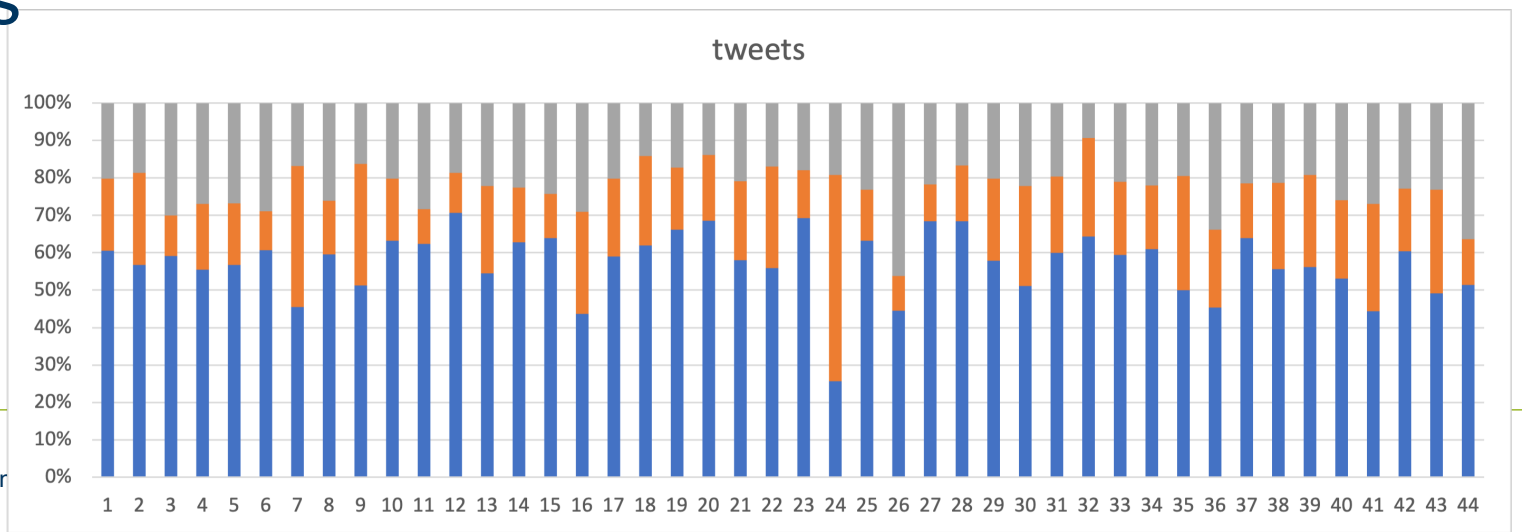
Personal pronouns

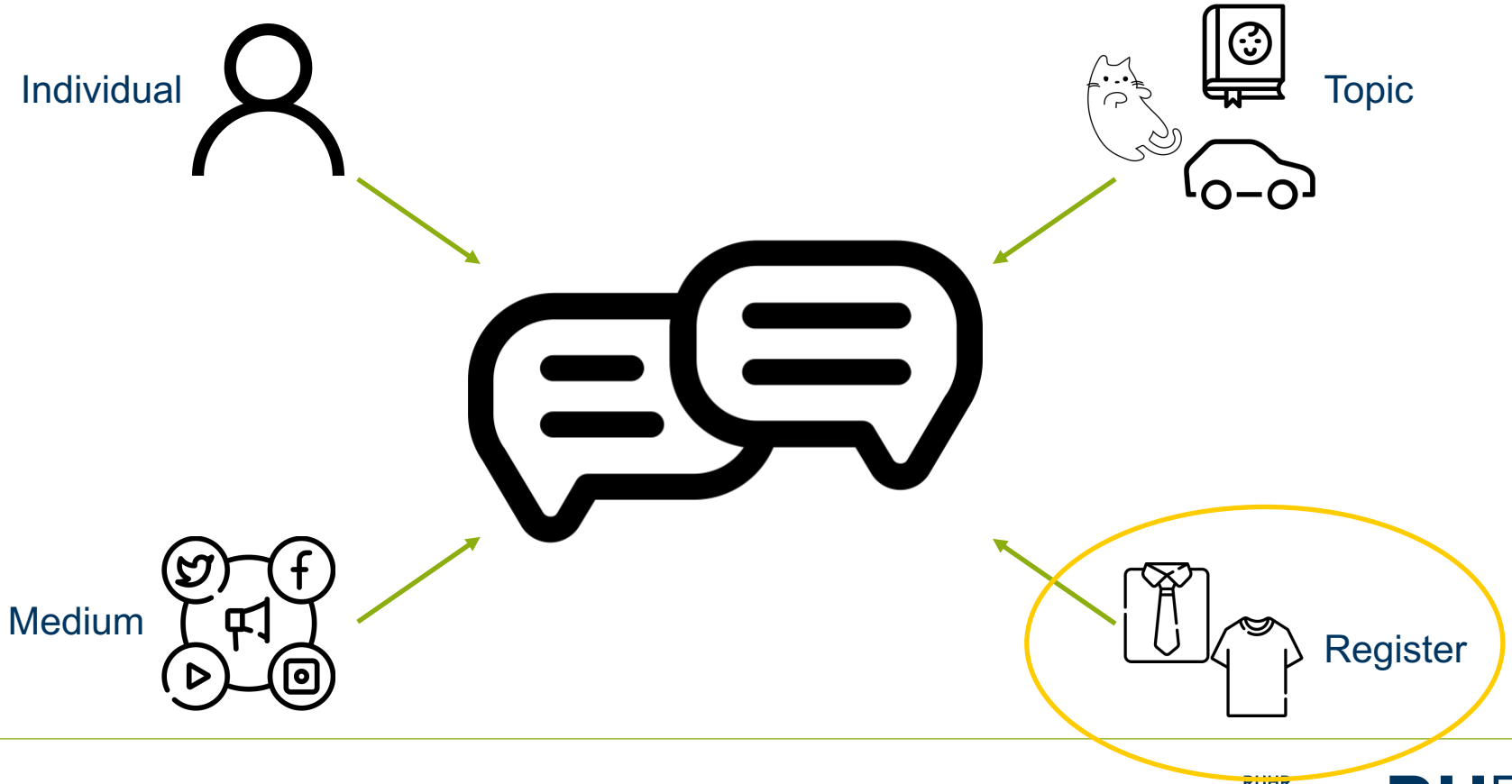
- Potsdam Commentary Corpus (German news): 88% third person pronouns
- Twitter / blogs:



Pronouns


blog posts
1st
2nd
3rd
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44


tweets
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44


Individual
Topic
Medium
Register

Register:

The Medium is not the Message

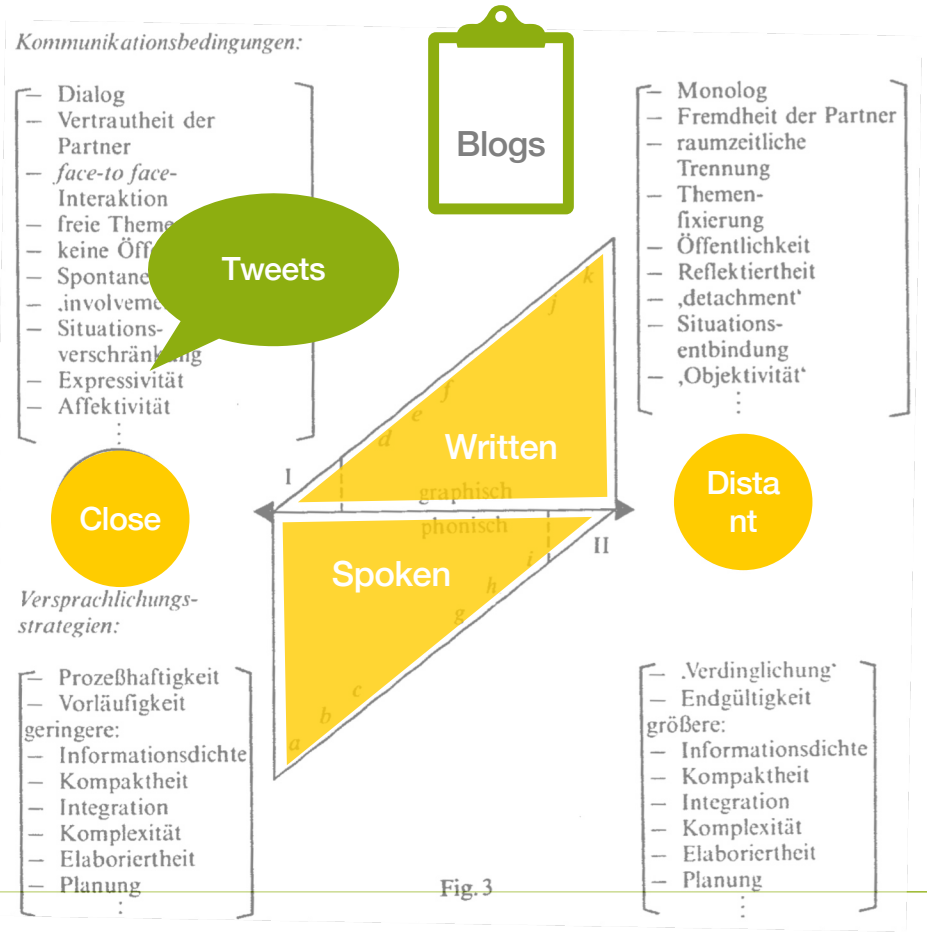


Fig. 3

Koch/Österreicher (1985)

CMC and register

- Register ≙ situational context of communication (Biber & Conrad 2005)
- What influences register? (Biber & Conrad 2019)
  - Setting and channel of communication
  - Discourse participants and their relationships
  - Purpose of communication
- Is there a social media register?
  - Tagliamonte 2016: ‘different CMC registers’, such as instant messaging, texting on phones, and email
  - Biber & Egbert 2016: ‘new internet registers, such as blogs, Facebook/Twitter posts, and email messages’

…But what about varying use of style in one social medium?

CMC and register

(1) ‘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]

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

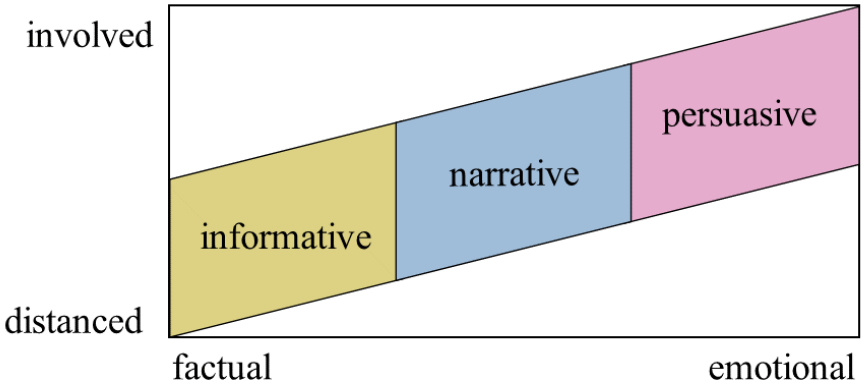
‘Boys are asleep . Husband is out … What about you all ? [URL]’

[tweets-2191]

CMC registers

- Defining registers based on non-linguistic factors (Biber & Conrad 2019):
  - Purpose
  - Topic
  - Reader involvement
  - Involvement of author

Identify three registers based on situational parameters



CMC register definitions

	Informative	***Narrative***	***Persuasive***
purpose	Passing on information, showing expertise, drawing attention to a topic, notifications, transfer of knowledge, announcements	Reporting everyday life and experiences, authenticity, relatability, sharing experiences, diary	Influencing readers, activism, politics, change, awareness, positioning, promotion
Involvement of author	Low, rarely says ‘I’ or ‘me’, personal position not expressed	Medium to high, depending of content and aspired level of relatability	High, speaks from own point of view, position visible and distinctly expressed, says ‘I’ or ‘me’, argumentation on a personal level
		…	

Register annotation

(1) ‘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**

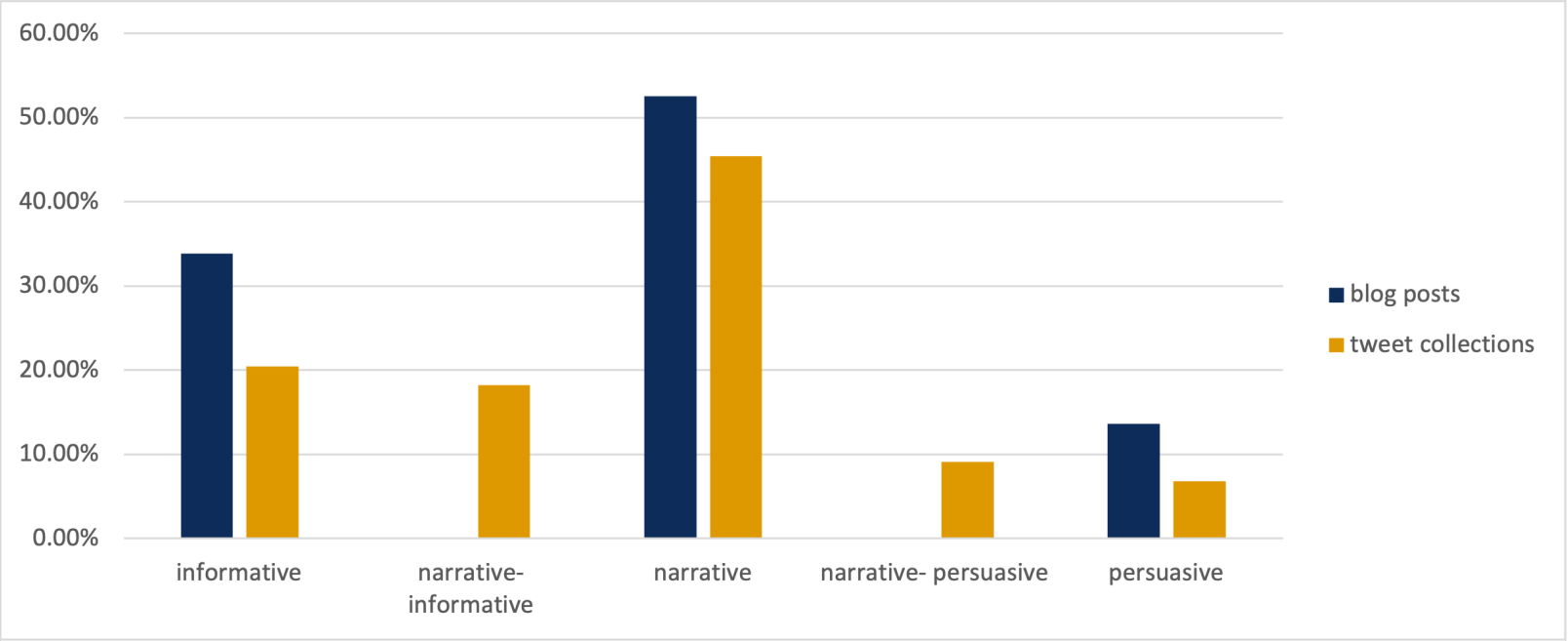
(2) ‘Jungs schlafen . Gatte ist aus ... Und ihr so ? [URL]’

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

[tweets-2191]

👉 **narrative register**

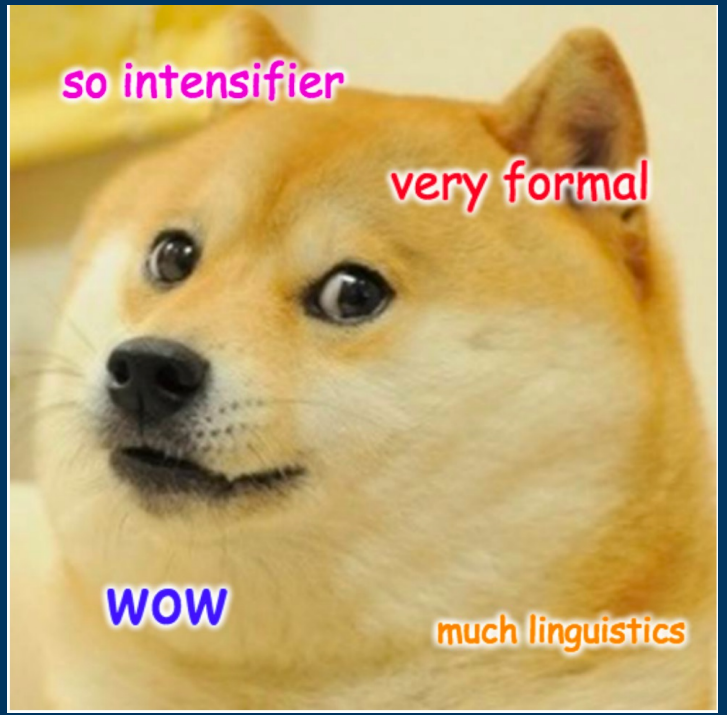
Register in TwiBloCoP



CMC register and linguistic phenomena

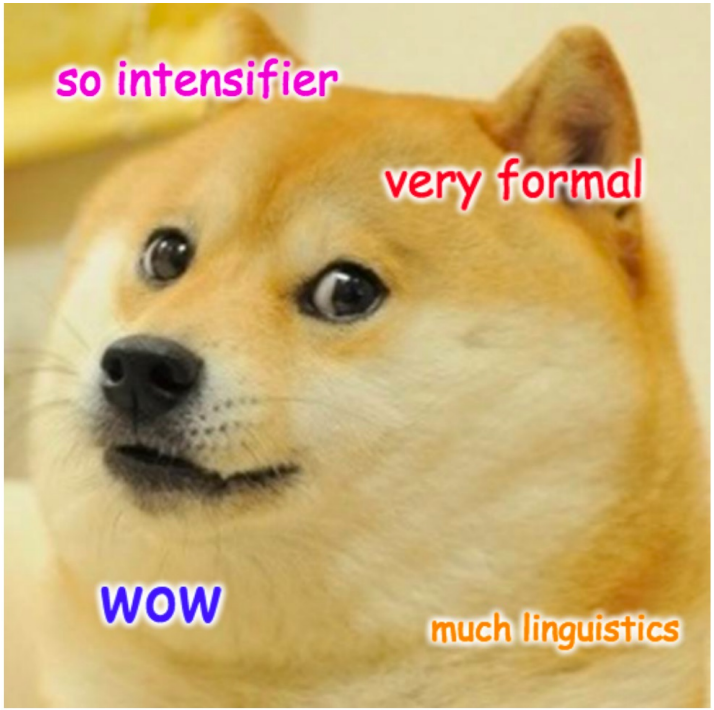
- Analyzed modal and intensifying particles in TwiBloCoP
- Cross-reference linguistic features and medium + register
- Found that medium and register both (independently) influence the use of these particles

Intensifiers



Intensifiers

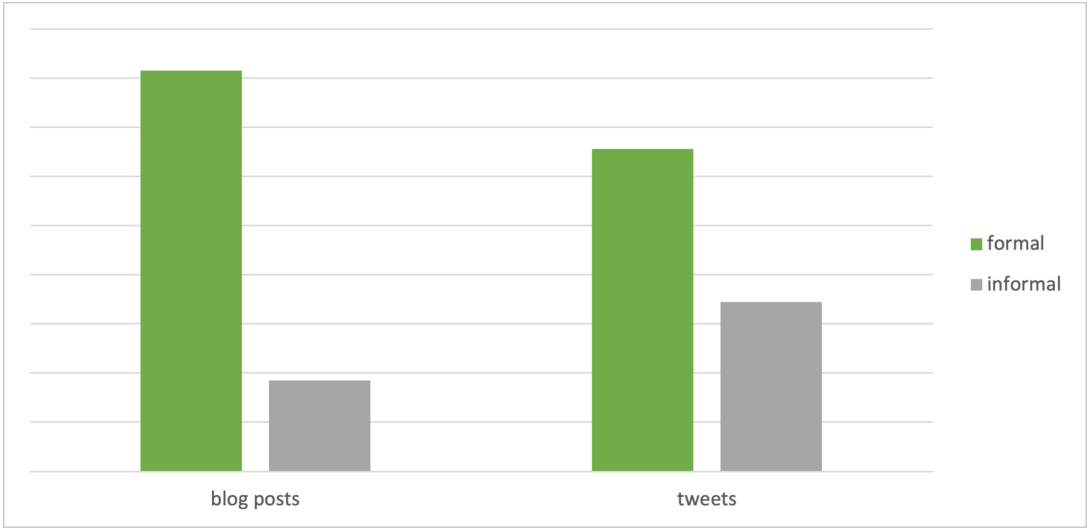
- Intensifiers “add intensity” to an utterance or property
- 2 contributions:
  - Narrow semantic: heightened degree
  - Not-at-issue: expressive value
- 37.2% of intensifiable adjective instances in fact have an intensifier in spoken German (Stratton 2020)
- Large variability across age groups and individuals



Intensifiers in CMC

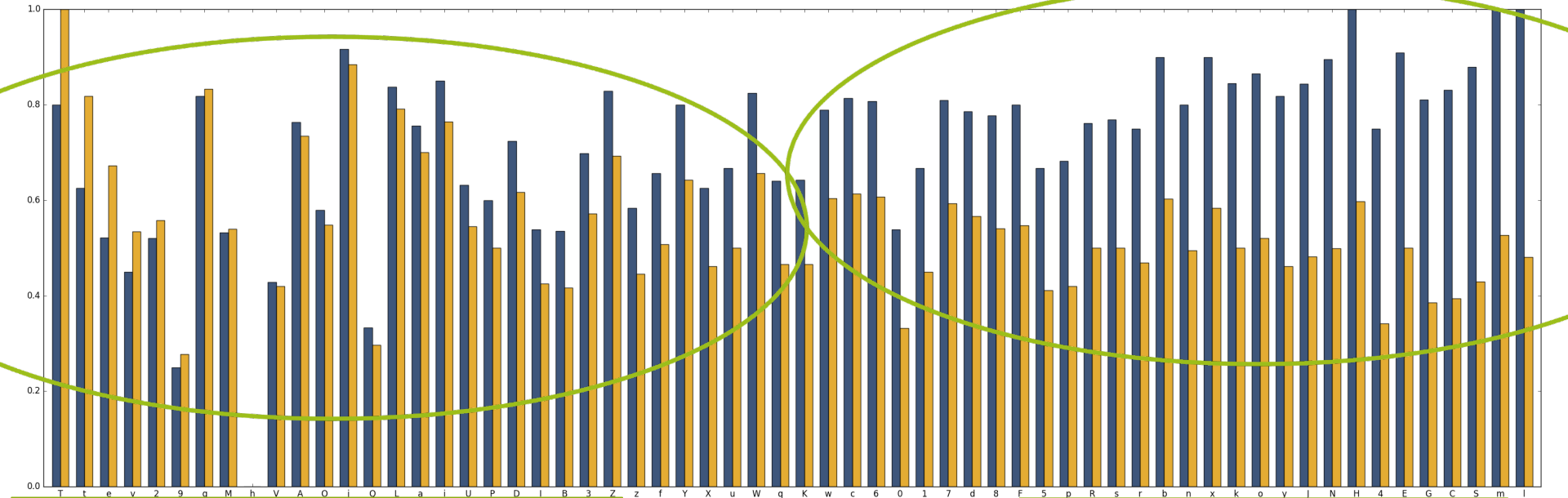
- Variability of intensifiers has long been reported
- Relative frequency of established (‘really, absolutely, very’) and more informal (‘totally, crazy, extremely, …’) intensifiers

- Manual annotation of intensifiers
- Top 5: *so, sehr, ganz, gar, wirklich*
- Most frequent: ‘so’ (occurs 3552 times)



Individual variation in intensifier usage

Percentage of formal intensifiers (*wirklich, sehr, absolut* = *really, very, absolutely*)

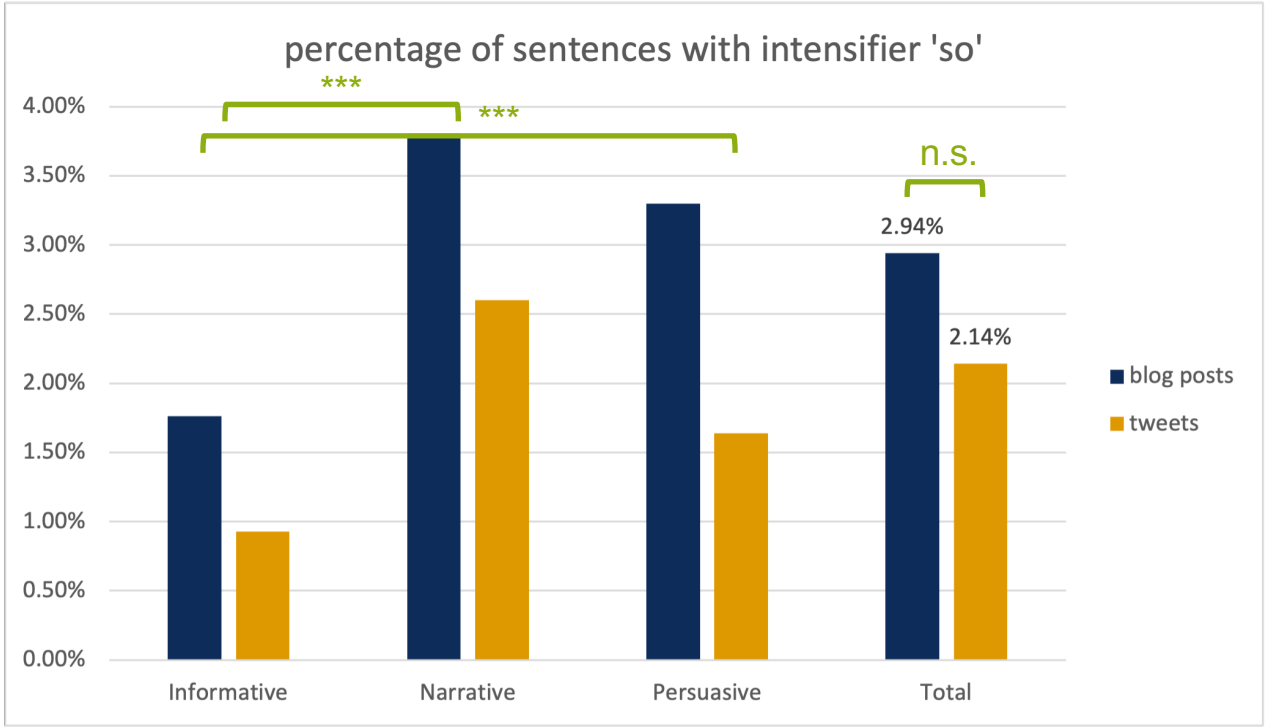


~ same usage in blogs/tweets

~ more formal in blogs

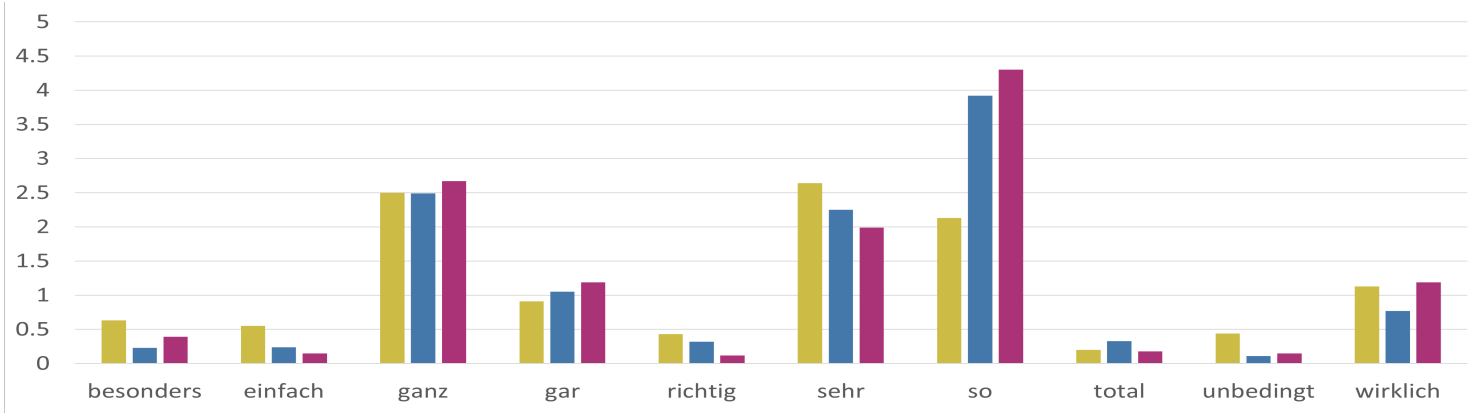
RUHR UNIVERSITÄT BOCHUM RUB

Intensifier usage varies by medium and register

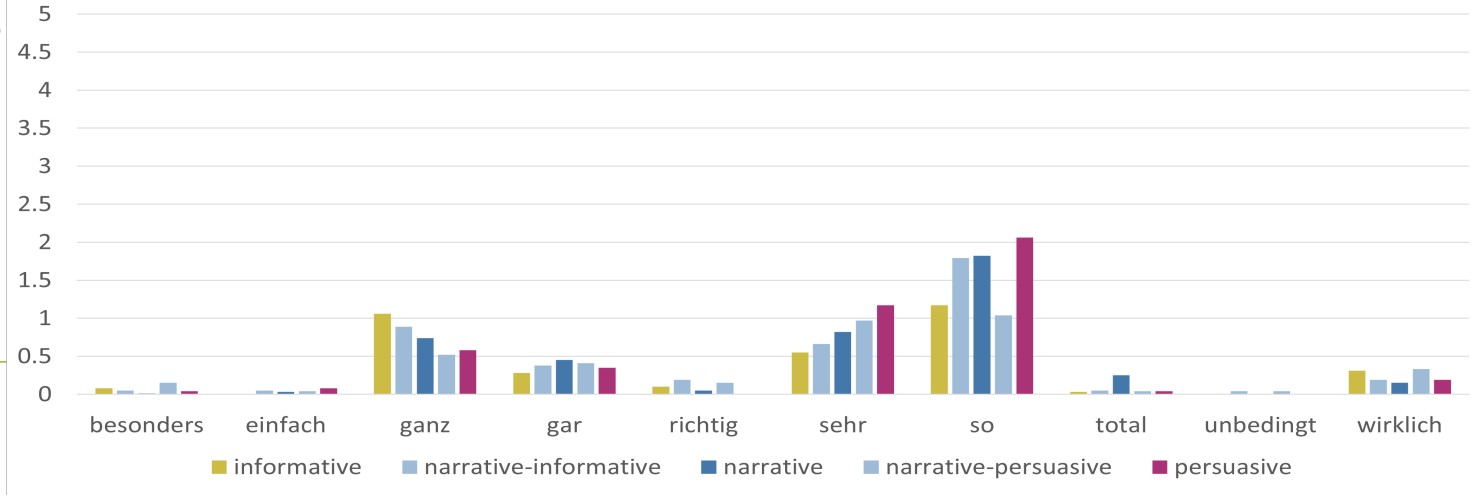


Count of intensifiers

blogs



tweets



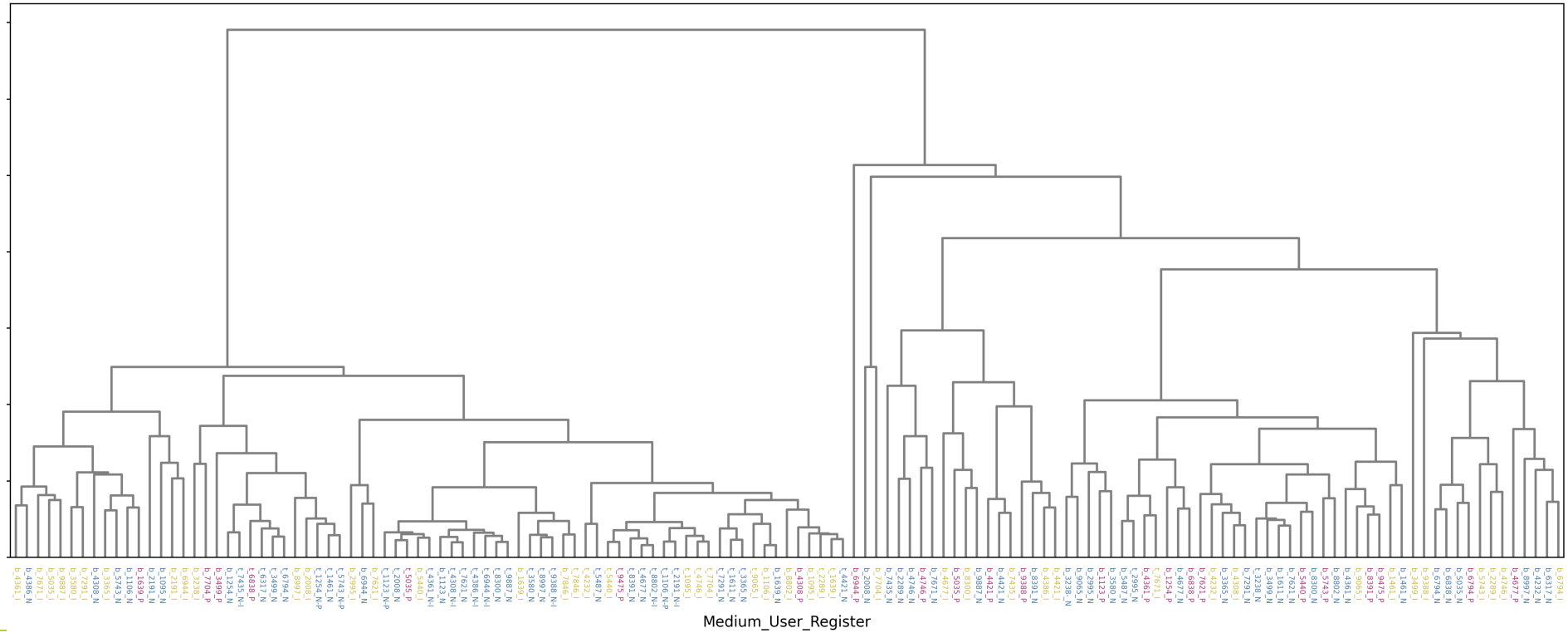
Clustering the corpus by linguistic features

Which documents are more similar to each other? (Author / Medium / Register)

- Agglomerative clustering using scikit-learn package
- For each user:
  - divide documents by medium and register
  - determine relative frequency of top 10 modal particles and intensifiers

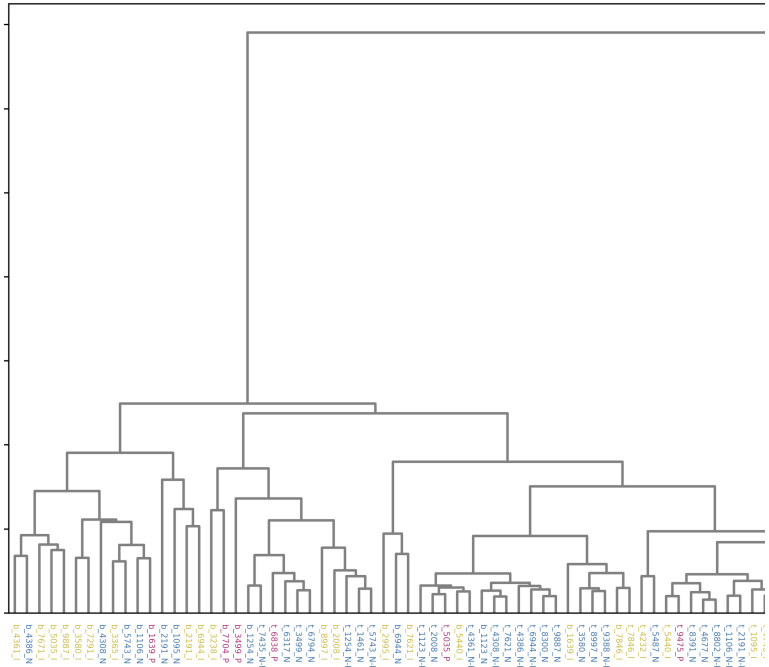
document	vector
b_1095_I	[0, 0, 1.13, 0.38, 0.38, 1.13, 1.5, 0, 0, 0, 0, 0, 0, 0, 0, 0.39, 0.39, 0.78, 0, 0]
b_1095_N	[0, 0, 0, 0, 0, 1.75, 0, 0, 0, 0, 0, 3.57, 0, 0, 1.79, 0, 0, 0, 1.79, 0]
t_1095_I	[0, 0, 1.59, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1.64, 0, 0, 0]

Result of linguistic clustering



Result of linguistic clustering

- Are register or social medium clustered more closely?

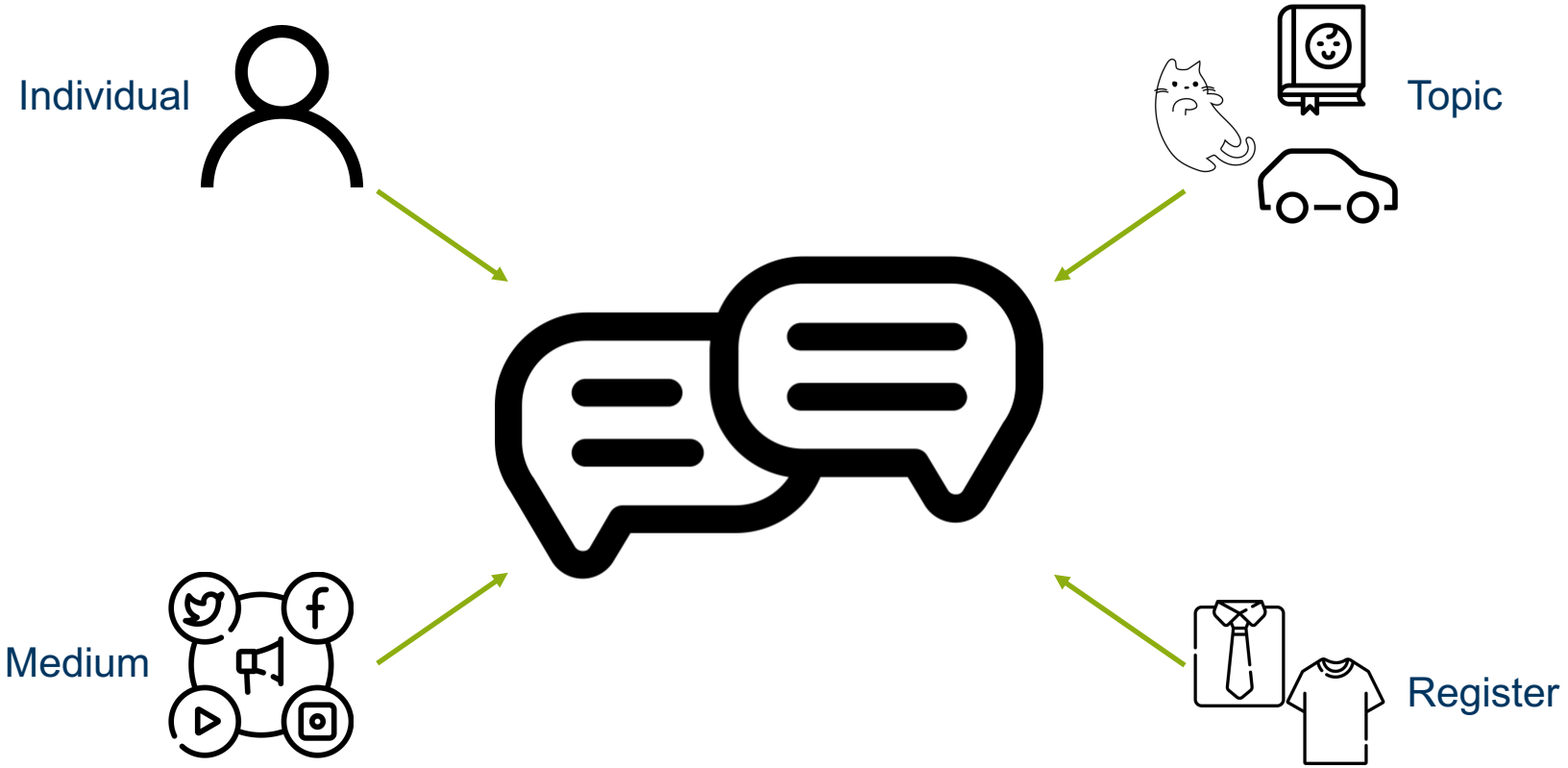


- V-measure (quality of grouping):
  - Medium V = 0.2246
  - Register V = 0.0839
- Homogeneity of clusters:
  - Medium: 0.5449
  - Register: 0.1419

Clustering conclusions

- What effects of the used medium and register can be differentiated from individual authors' variability?
  - Use of intensifiers differs depending on medium and register
- How well does bottom-up clustering based on specific linguistic features capture authors' usage of different registers and media?
  - Modal and intensifying particles can be used as feature for clustering
  - Clustering corresponds more closely to medium distinction

Summary


Individual
Topic
Medium
Register

Individual variation in CMC

- No “CMC register”
- Must control for topic, medium, register, individual properties of authors
- All can have independent effect on linguistic phenomena
- Large corpora cause huge effort to
  - Collect
  - Preprocess
  - Annotate
  - Analyze
- Share data
  - Verification
  - Sustainability

Bochum CMC Corpora

- TwiBloCoP

- XXL: eXXtra-Large German Twitter corpus
  - 2014–2023
  - Majority of German tweets
  - No topic/keyword based filtering
  - ~ 2 billion tweets
  - Access: German National Library (soon) or contact me

- ChrisTof
- PARADISE

Save the Date!

Sustainable Archiving, Indexing, and Distribution of Dynamic Data from Social Media – Twitter and Beyond.

German National Library, Frankfurt

Conference: March 19–20, 2024

Data Sprint: March 21–22, 2024

More Info, Call for Abstracts, Registration: Stay tuned.

THANK YOU!

Tatjana Scheffler

tatjana.scheffler@rub.de

Image references:

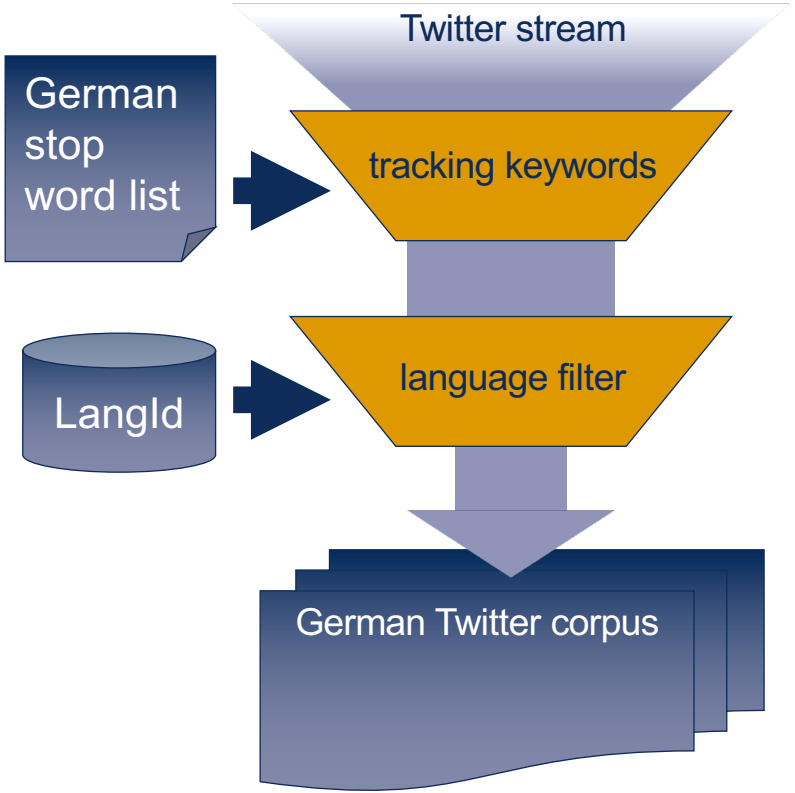
- Title: Lewis Ogden, https://www.flickr.com/photos/bitsfrombytes/43617178595
- text by Alice Design from Noun Project (CC BY 3.0)
- Cat by lizbeth from Noun Project (CC BY 3.0)
- kids by Vectorstall from Noun Project (CC BY 3.0)
- Car by Rainbow Designs from Noun Project (CC BY 3.0)
- User, social media, 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.
- Hannah Seemann and Tatjana Scheffler. Differentiating Social Media Texts via Clustering. In: Proceedings of the CHR Conference 2022, pp. 177–188. 2022. https://ceur-ws.org/Vol-3290/short_paper5903.pdf
- Tatjana Scheffler, Michael Richter, Roland van Hout. Tracing and classifying German intensifiers via information theory. Language Sciences 96. 2023. https://doi.org/10.1016/j.langsci.2022.101535 Repository location: https://osf.io/69x8b/

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 tweets in total
- Access: German National Library (soon) or contact me



CMC register definitions

	Informative	***Narrative***	***Persuasive***
topics	Milestones in personal life, competitions, reviews, recipes, nutrition, crafting	Everyday life, holidays, career, health, nutrition, personal life	Parenting, nutrition, product placement, politics, gender, education, sustainability
Interactivity, reader involvement	Low, rarely addresses reader, no engagement, no call to action, no reaction needed / expected	Optional, can be medium to high, depending on format/level of community	High, expected, might be provocative, direct address, high level of emotionalization