RUHR UNIVERSITÄT BOCHUM RUB

Computational Linguistic Approaches to Digital Conversations: the Case of Intensifiers

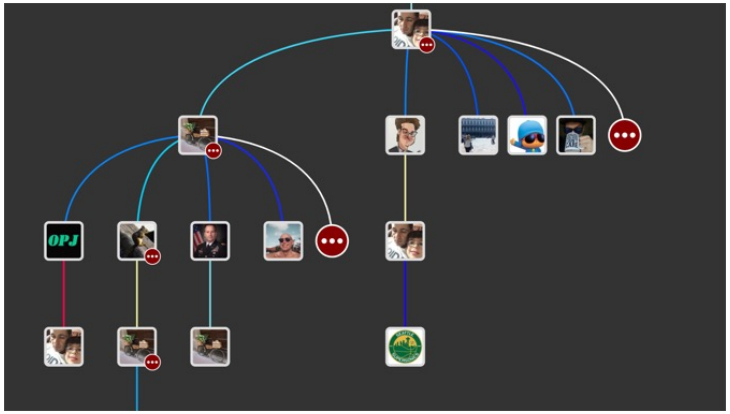
University of Gothenburg

Tatjana Scheffler
tatjana.scheffler@rub.de
@tschfflr@fediscience.org

March 17, 2023

Digital Forensic Linguistics

- Basic research
  - Linguistics of lying and deception
  - Digital linguistics
  - Discourse structure of online conversations
  - Linguistic variability
  - Non-literal meaning
- Applications
  - Disinformation detection
  - Hate speech detection
  - Authorship analysis



digital linguistic data

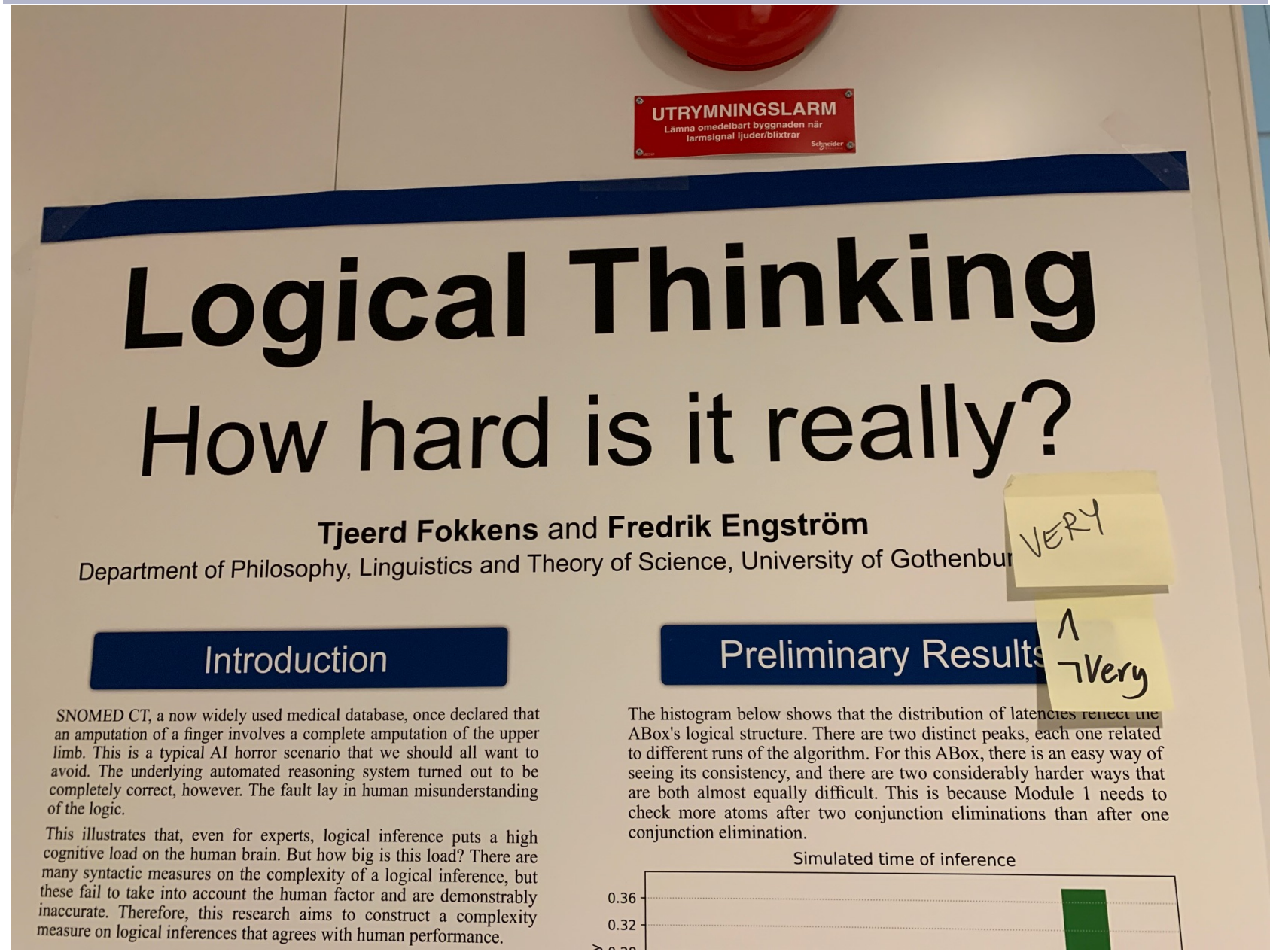
Team

Intensifiers (really, so, very)

- 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:

(1) This seal is extremely fat.

(2) This seal is ultra fat.


UTRYMNINGSLARM
Lämna omedelbart byggnaden när larmsignal ljuder/blixtrar
Schneider
Logical Thinking
How hard is it really?
Tjeerd Fokkens and Fredrik Engström
Department of Philosophy, Linguistics and Theory of Science, University of Gothenbur
VERY
∧
¬Very
Introduction
SNOMED CT, a now widely used medical database, once declared that an amputation of a finger involves a complete amputation of the upper limb. This is a typical AI horror scenario that we should all want to avoid. The underlying automated reasoning system turned out to be completely correct, however. The fault lay in human misunderstanding of the logic.
This illustrates that, even for experts, logical inference puts a high cognitive load on the human brain. But how big is this load? There are many syntactic measures on the complexity of a logical inference, but these fail to take into account the human factor and are demonstrably inaccurate. Therefore, this research aims to construct a complexity measure on logical inferences that agrees with human performance.
Preliminary Results
The histogram below shows that the distribution of latencies reflect the ABox's logical structure. There are two distinct peaks, each one related to different runs of the algorithm. For this ABox, there is an easy way of seeing its consistency, and there are two considerably harder ways that are both almost equally difficult. This is because Module 1 needs to check more atoms after two conjunction eliminations than after one conjunction elimination.
Simulated time of inference
0.36
0.32

Joint work with…

- Michael Richter (Leipzig) and Roeland van Hout (Radboud U.)
- Hannah Seemann, Imge Yüzüncüoglu (Bochum)
- Lesley-Ann Kern (Bochum, now Marburg)
- Tariq Youssef, Nathanael Philipp (Leipzig)

https://imgs.xkcd.com/static/tour_challenge.png

Linguistic variability in large digital data

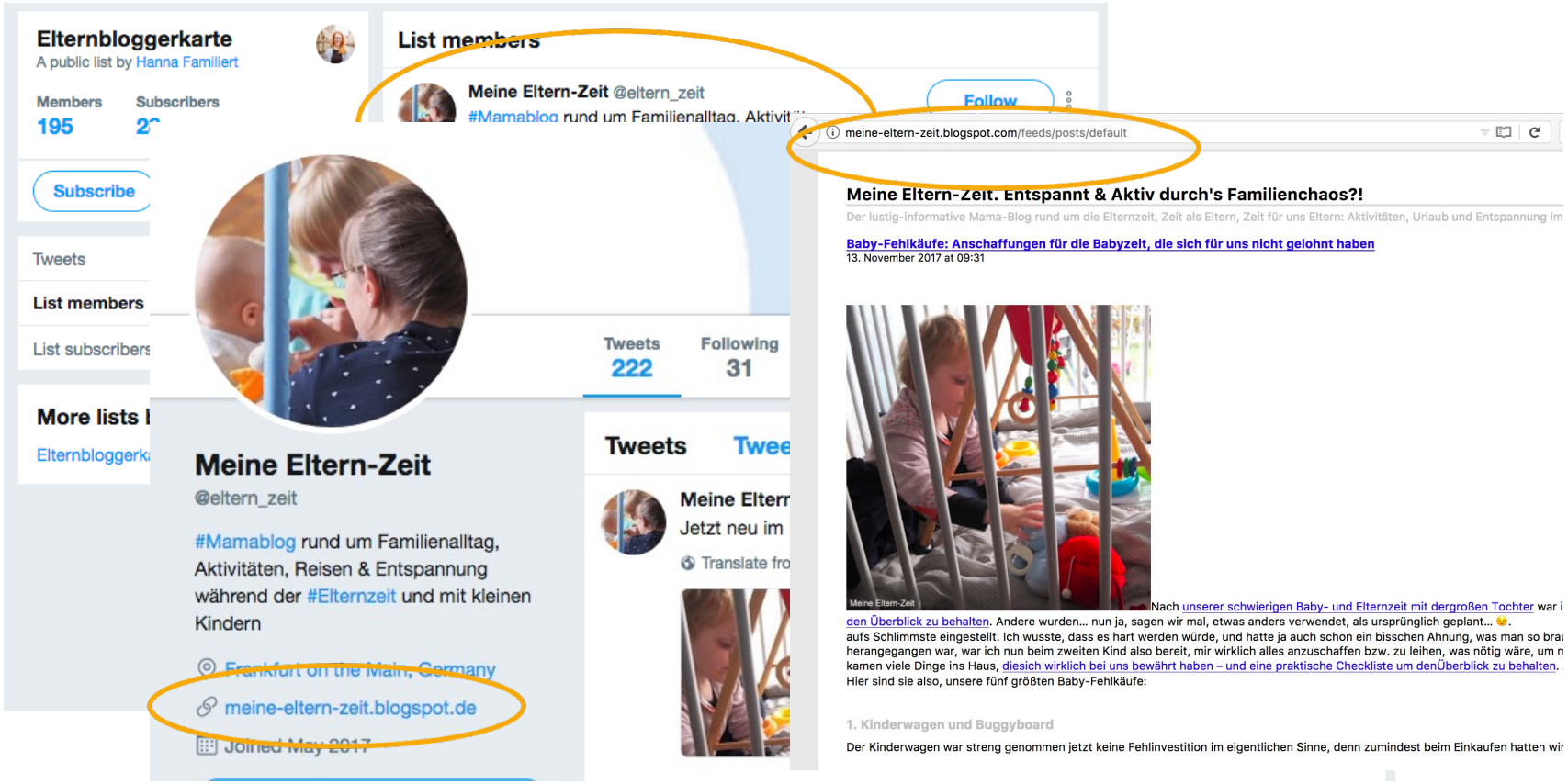
(Scheffler/Richter/van Hout 2023, Scheffler/Seemann/Kern 2022)

Variability on social media

- Social media = informal language in written form
- Phenomena typical of spoken language
- Speakers' language use differs between media, communities, etc.
- Discourse level rarely studied

=> Create a corpus of different media with the same users

Creating a cross-channel corpus



TwiBloCoP (Twitter+Blog Corpus – Parenting)

- Topic: family life and parenting
- Collection: Oct 2016–Feb 2017
- Explicit retrieval of author consent (opt-out)
- Manual pseudonymization of personal names, locations, and other identifying details:

„Clean sidewalks! In **XXXXX** they just flatten the snow and put insanse amounts of gravel (or whatever that‘s called) on top.“

→

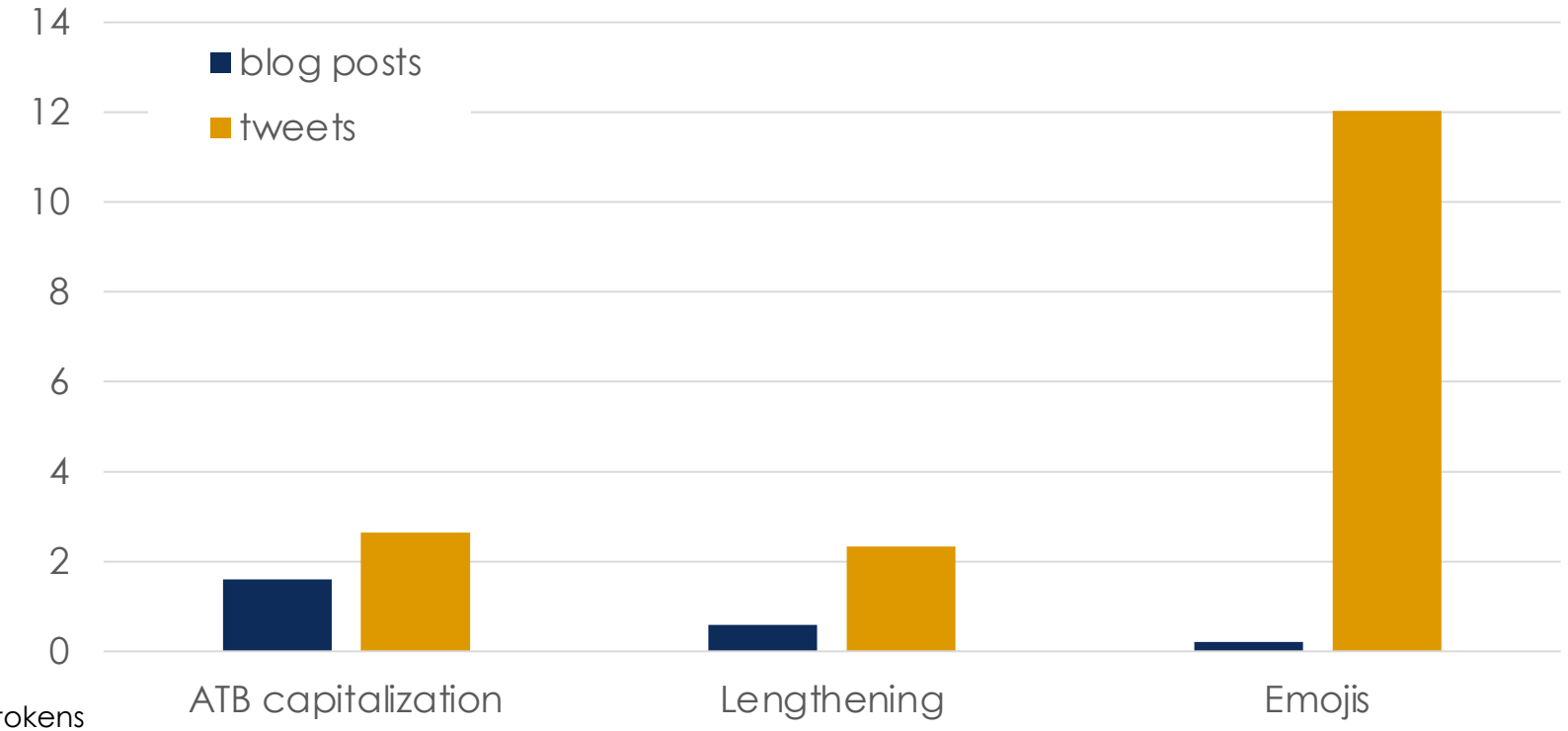
„Clean sidewalks! In [LOC] they just flatten the snow and put insanse amounts of gravel (or whatever that‘s called) on top.“

Overview: Cross channel corpus

	Blog posts	Tweets	PCC (news)
users	44	44	
items	468	81,440	
tokens	361,117	1,170,888	
type/token ratio (avg.)	0.28	0.22	0.54
word length (chars.)	4.68	4.85	6.36

- TTR over first 1000 tokens (per user)
- complexity measures indicate similarity to spoken data

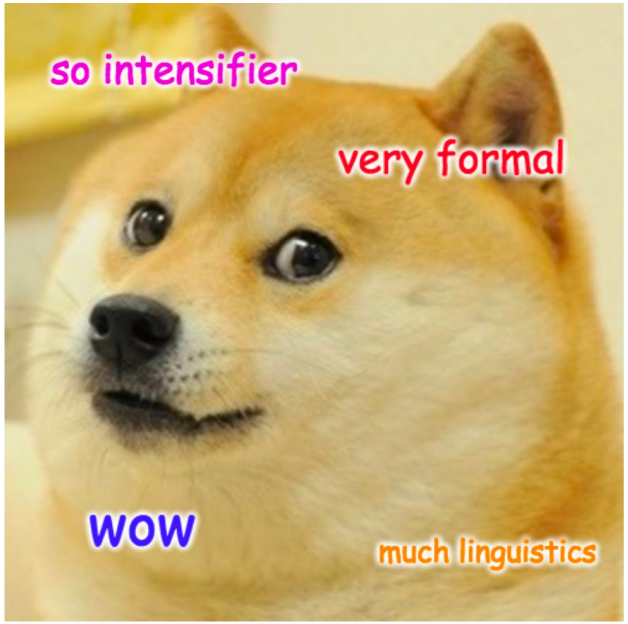
Social media items

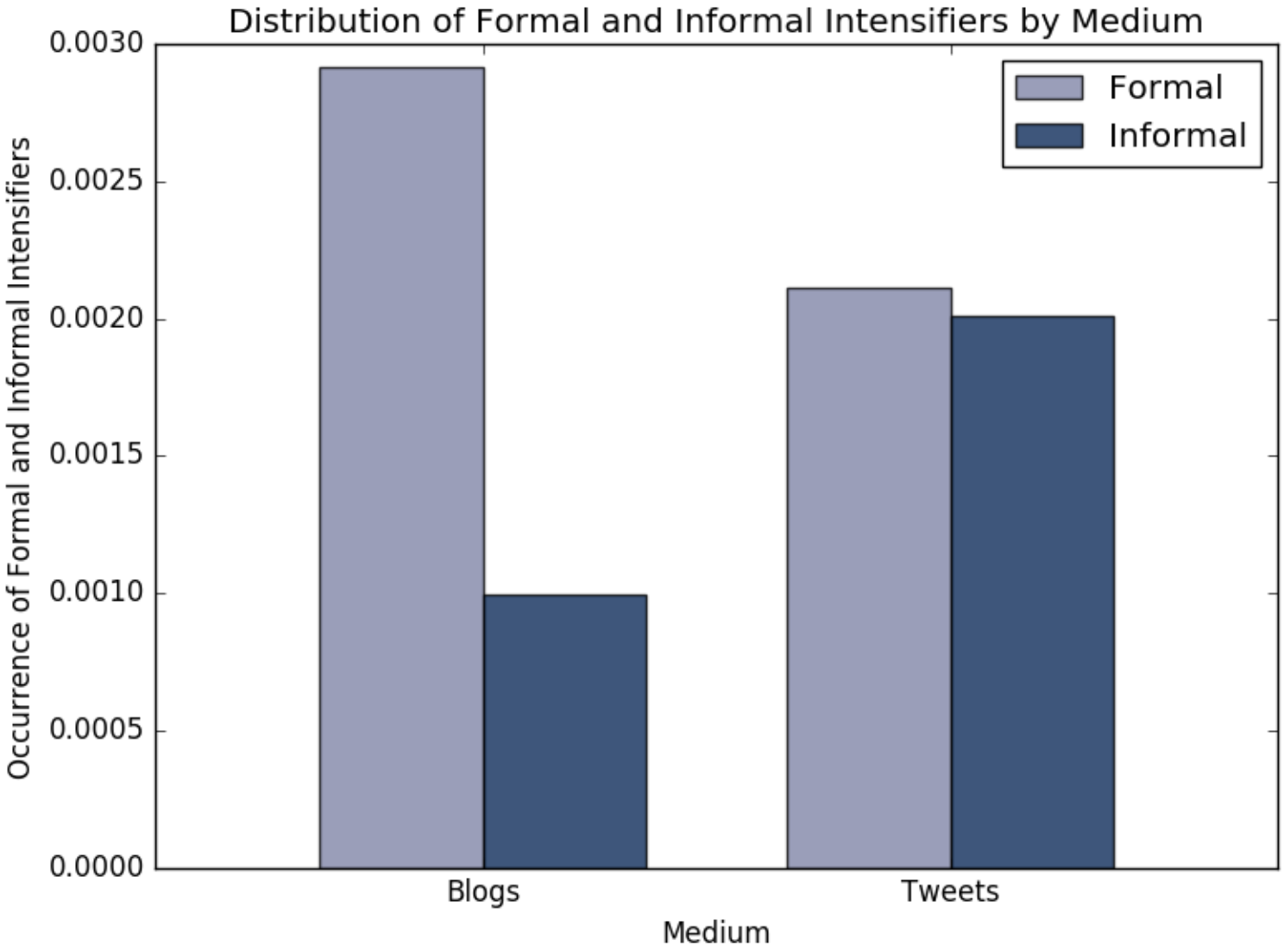


Per 1000 tokens

Intensifiers in social media

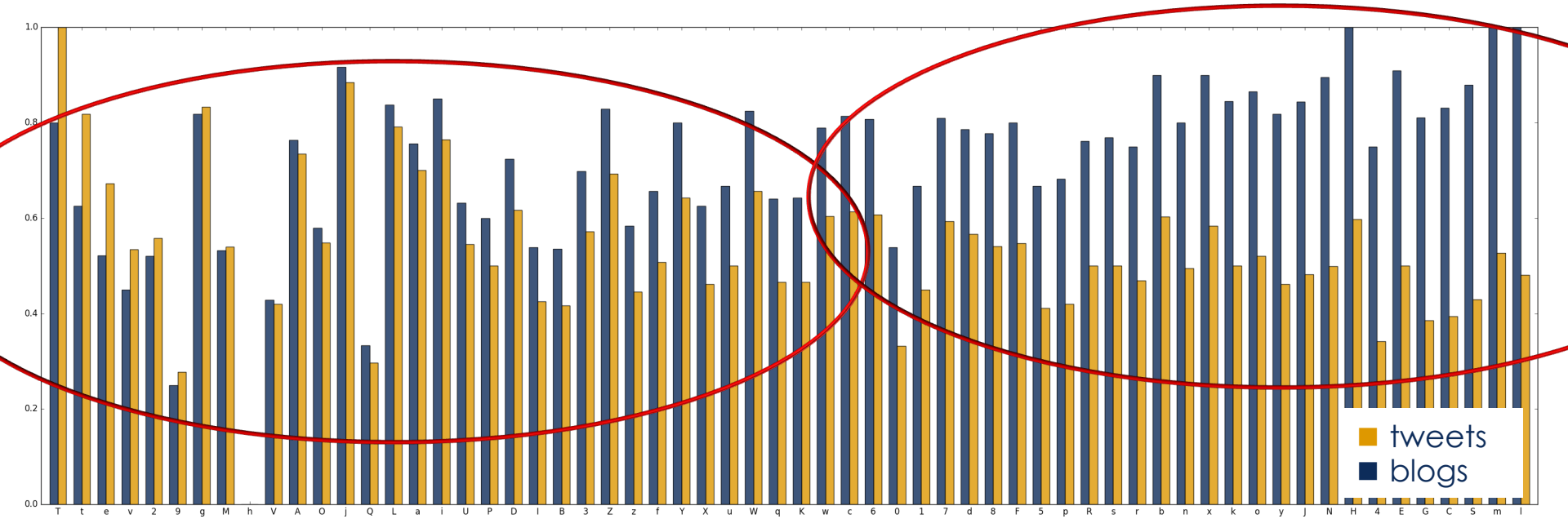
- Use of intensifiers is associated with colloquial registers (speech)
- Formal vs. informal: speaker's choice
- Intensifiers are equally frequent in both media (3x as frequent as in news text)
- Formal: 'wirklich', 'sehr', 'absolut'
- Informal: 'echt', 'krass', 'extrem', 'voll', 'völlig', 'total', 'ordentlich', 'sau'




Distribution of Formal and Informal Intensifiers by Medium
Occurrence of Formal and Informal Intensifiers
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
Formal
Informal
Blogs
Tweets
Medium

Intensifiers, individually

- percentage of formal intensifiers among all intensifiers



~ same usage in blogs/tweets

~ more formal in blogs

Intensifiers in digital media

- frequent use in social media
- on aggregate more formal in blogs, less so in tweets
- some individual differences: Some users employ mostly formal intensifiers in blogs, some behave similar to the way they behave on Twitter

=> Model intra-speaker variability

Intensifiers and Information Theory

What actually is an intensifier?

- 2 contributions:
  - Narrow semantic: heightened degree
  - Not-at-issue: expressive value
- What is the difference between the many intensifiers in a language?
- Can information-theoretic notions explain the choice of intensifier?
- What determines the order of stacked intensifiers?

Intensifiers in German

- Large and changing set of words (Claudi 2006; Scheffler et al. 2023: 124 frequent intensifiers)
- They differ in expressivity: how much intensification is expressed
- Frequent, old intensifiers get semantically bleached (weaker), but combine with more different adjectives and intensifiers
- New intensifiers are more surprising and stronger

Intensifier	Approximate translation	Tokens
so	so	33783
sehr	very	10403
echt	really	9861
voll	fully	6855
ganz	completely	5935
einfach	simply	5566
wirklich	really	3744
total	totally	3163
richtig	right	2609
mega	mega	1264
schön	beautifully	1001
verdammt	damn	970
super	super	847
extrem	extremely	648
unglaublich	unbelievably	603
völlig	totally	596
gut	good	572
sau	(female) pig	565
sowas von	so	537
absolut	absolutely	324
vollkommen	totally	247
schwer	hard	202
scheiße	shit	192
komplett	completely	189
krass	crass	186
unfassbar	unbelievably	184
geil	horny	172
gar	even	160
besonders	especially	147
arg	very	141
hammer	hammer	139
übelst	awfully	138
ur	very	136
über	over	133
fucking	fucking	128
wahnsinnig	insanely	123
scheiß	shit	112
äußerst	extremely	103
unendlich	infinitely	94
furchtbar	terribly	87

Intensifier	Approximate translation	Tokens
doppelt	doubly	82
dermaßen	so	75
unheimlich	eerily	70
perfekt	perfectly	69
unbeschreiblich	indescribably	68
gold	gold	66
schrecklich	awfully	63
übertrieben	exaggerated	61
zutiefst	profoundly	60
übel	awfully	56
ultra	ultra	51
erstaunlich	amazingly	51
hart	hard	49
stark	strong	49
höchst	highest	48
irre	crazy	45
derbe	crudely	45
hoch	high	44
wunder	miraculously	37
genial	ingeniously	37
traumhaft	dreamlike	35
wunderbar	wonderfully	35
extra	extra	33
ernsthaft	seriously	32
tierisch	animal-like	31
arsch	ass	31
tief	deep	30
extremst	most extreme	30
abartig	degraded	29
reichlich	plenty	29
wunderschön	miraculously beautiful	28
doll	much	28
very	very	27
schlimm	badly	27
top	top	26
heftig	fiercely	26
krank	sick	26
unnormal	abnormal	25
enorm	enormously	25
brutal	brutally	24
mies	lousy	24

Intensifier stacking

(1) Das ist doch hammer mega geil

'That is (particle) hammer mega awesome'

(2) a. Frankfurt ist so arsch weit

'Frankfurt is so damn far'

b. ? Frankfurt ist arsch so weit

? 'Frankfurt is damn so far'

- What is the reason for intensifier stacking?
- What explains the strong preferences for intensifier ordering?

→ information theory

Data collection

- Conversation threads from Twitter (Scheffler, 2014): 6 mio tweets
- Extract possible predicative phrases:

  Sie ist mutig

  PPER VAFIN (6) (5) (4) (3) (2) (1) ADJD
- Semi-automatically select intensifiers from the adj. modifiers
- Re-extract intensified predicative phrases
- Final list: 124 frequent intensifiers (excl. downtoners; >5 occ.)
- 38 (30.6%) also occurred as an adjective in our data
- Focus on predicative phrases with 1-3 intensifiers

Information measures

- Local (paradigmatic) information (Shannon information content): IC_{local}

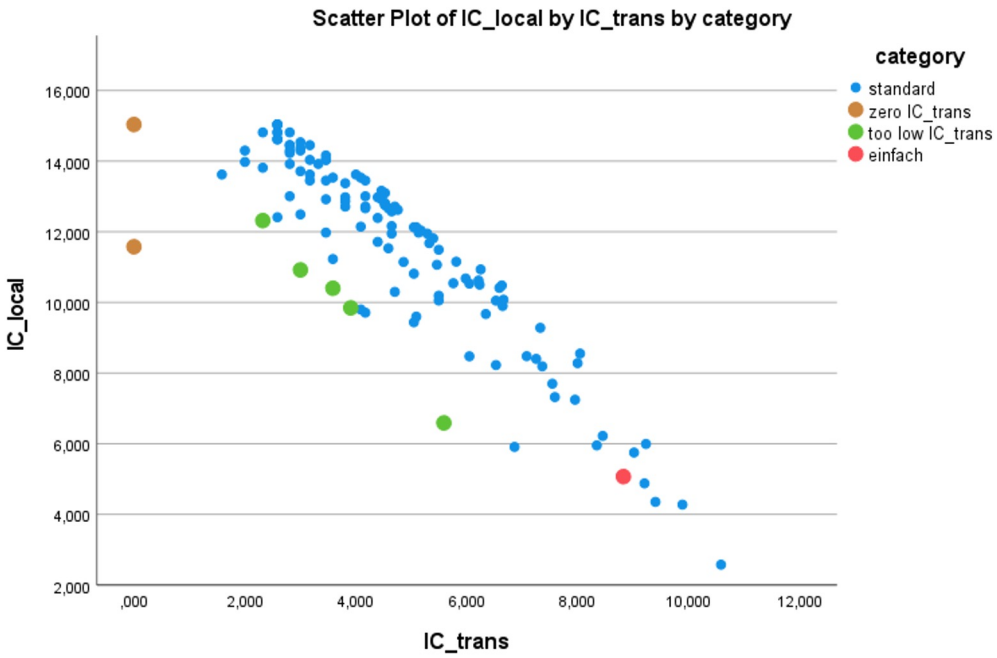
$$IC_{LOCAL} = -log_2 \frac{|w|}{|Intensifiers|})$$

- Contextual information content (Markov transition): IC_{trans}

$$IC_{TRANS}(w_t) = \overline{IC}(w_t) = -\frac{1}{n} \sum_{t=1}^{n} log_2 P(w_{t+1}|w_t)$$

Correlation of IC_{local} and IC_{trans}

- Correlation = -.916
- IC_{trans}= 0 : wunschlos ‘wishless’, gold ‘gold’
- Low IC_{trans} : eklig ‘disgusting’, geil ‘horny’, fett ‘fat’, großartig ‘great’, and mies ‘bad’
- Einfach ‘just’ : sentence adverb



Results: Intensifying intensifiers

- We hypothesize that intensifiers are further emphasized / strengthened by
  - stacking multiple intensifiers in one phrase
  - grapheme lengthening or capitalization
- 7492 out of 89358 intensified phrases (8.4%) contain stacked intensifiers
- Variants are also frequent

mega	9	18.9%
so	1	17.2%
sehr	2	3.5%
richtig	8	2.8%
total	7	2.5%
voll	4	2.3%
ganz	5	2.0%
schön	10	1.3%
wirklich	6	0.5%
echt	3	0.5%

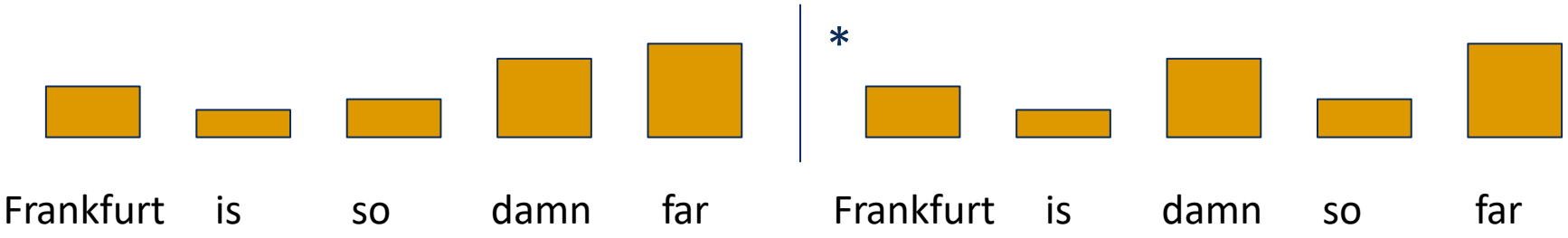
Intensifier stacking

- Several intensifiers in an AdjP increase the length of the phrase and may thus increase expressiveness (Bennett/Goodman 2018)
- (Richter/Van Hout 2020) observe for Dutch that an intensifier with a high use value, IC_{trans} , often precedes other highly unexpected and highly expressive intensifiers
- Less expressive intensifiers thus prepare the processor for other more unusual intensifiers. Their function as intensifier is also disambiguated by their position (between "vanilla" intensifier and adjective)

(2) Frankfurt ist so arsch weit
'Frankfurt is so damn far'

Uniform Information Density hypothesis

- Uniform distribution of information across a linguistic utterance (Levy & Jaeger 2007)
- Less informative words precede more informative ones; enable their predictability (Fenk-Oczlon 1989)
- Less expressive and surprising intensifiers thus prepare the processor for other more unusual intensifiers.



Results: Stacking order

- We predict increasing IC_{local} from left to right in intensifier stacks (least to most expressive)
- Out of 4858 pairs, 969 violate this expectation (20%)
- Most violations due to echt and wirklich 'really'

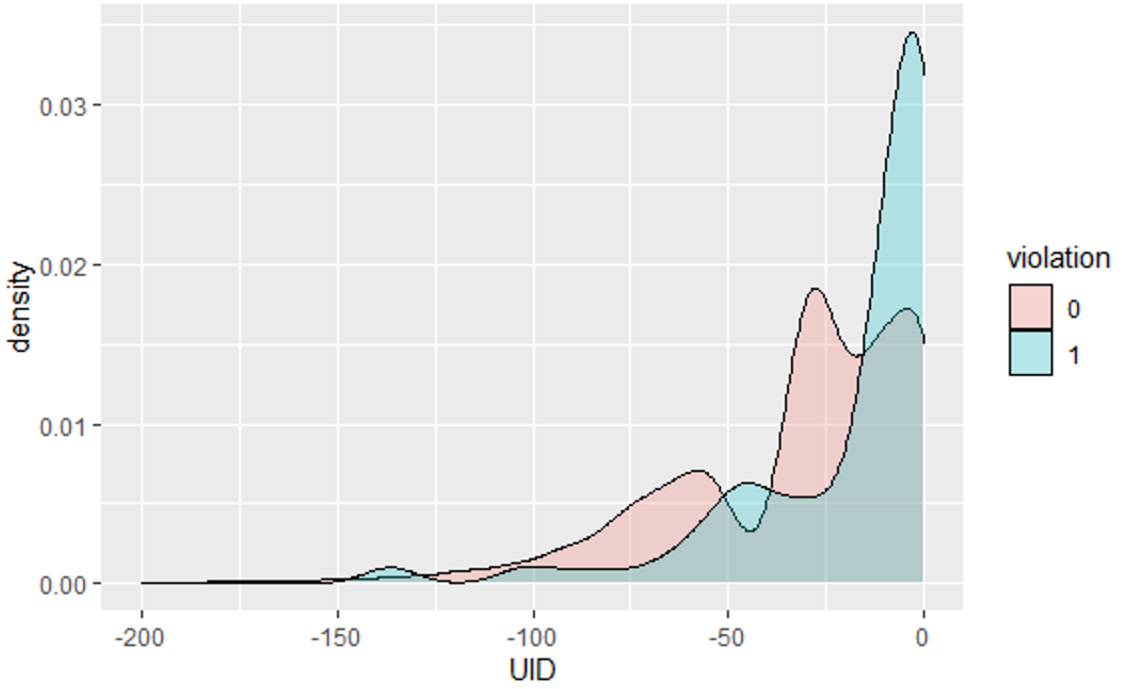
Remainder:

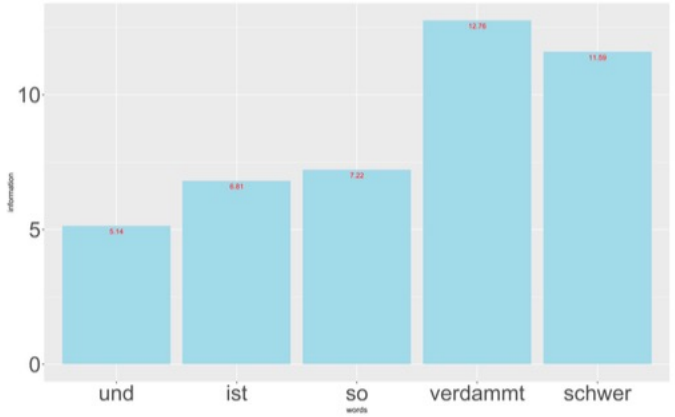
- 183 violations out of 3108 pairs (5.9%)
- Larger differences in information content should have a stronger effect on the stacking order
- Computed uniformity of information density for conforming and violating intensifier stacks

Results: Stacking order + information density

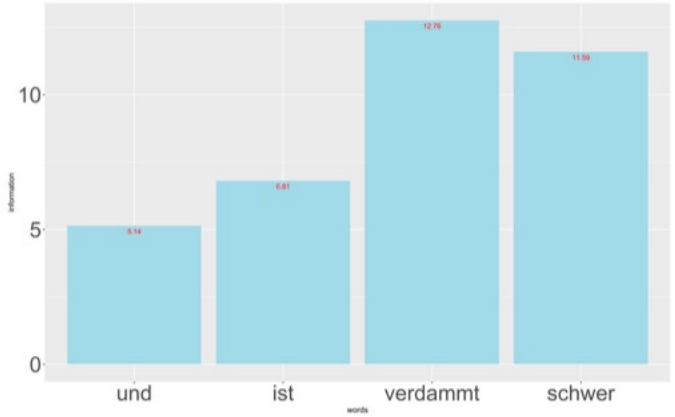
$$UID_{local} = -\frac{1}{N}\sum_{i=1}^{N}(id_{ij} - id_{ij-u})^2$$

- Violations have a different UID distribution
- The difference in information is very small in "violations"

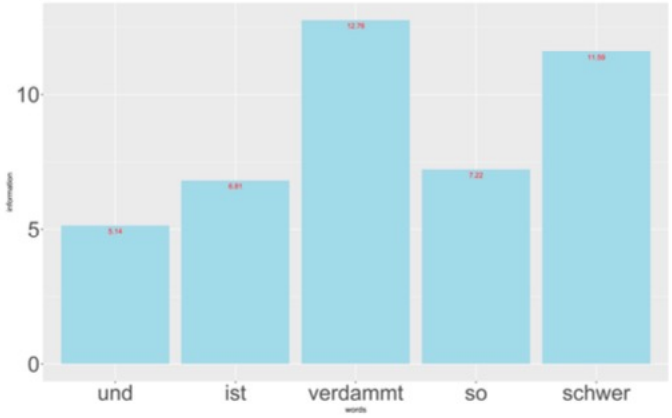




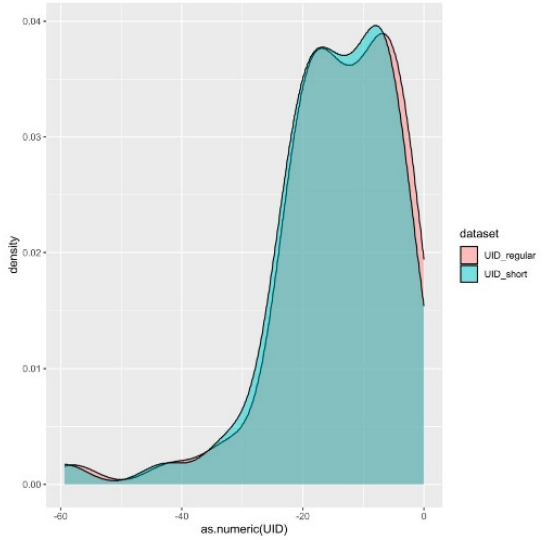
(a) Flow of *IC* in the phrase *und ist so verdammt schwer*.



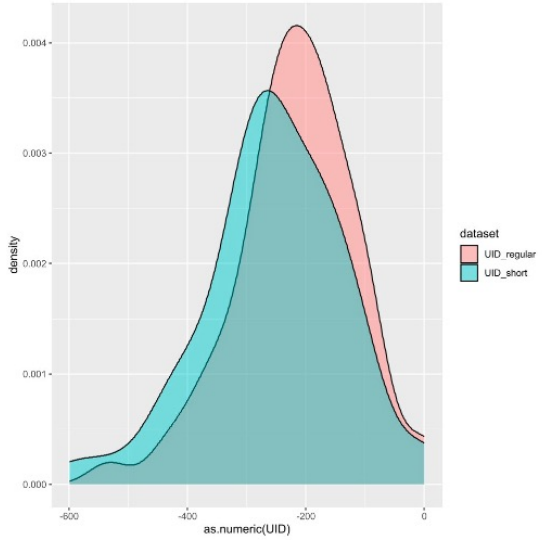
(b) Flow of *IC* when *so* is omitted.



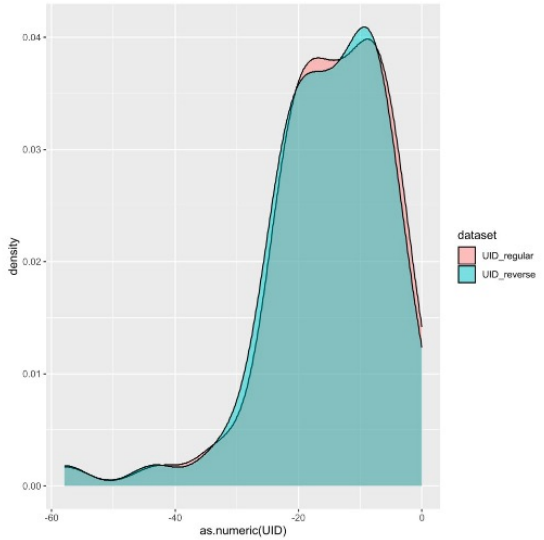
(c) Flow of *IC* when the order of *so* and *verdammt* is reversed.



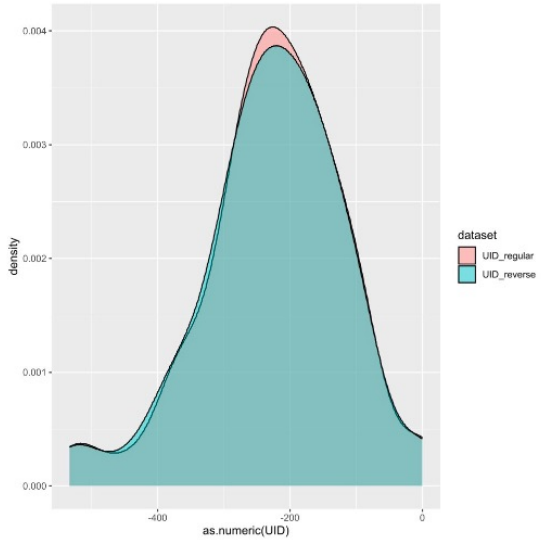
(a) Information density of *IC* in the blog data, compared to shortened stacks.



(b) Information density of *TCM* in the blog data, compared to short stacks.



(c) Information density of *IC* in the blog data, compared to reversed stacks.



(d) Information density of *TCM* in the blog data, compared to reversed stacks.

Intensifiers and Information Theory

- Intensifiers differ in their expressive value
- Newer/more informative intensifiers combine with fewer adjectives: extremely strong correlation between IC_{local} and the range of following adjectives IC_{trans} (-0.916)
- Intensifiers are chosen for their surprisal and further strengthened by lengthening, capitalization, and/or stacking
- "Vanilla" intensifiers precede more unusual ones and contribute to their interpretability
- Information-theoretic study of intensifiers allows us to identify "exceptions", such as ambiguous words (*einfach* 'simply', *echt/wirklich* 'really')

‘Einfach’, ‘echt’, and ‘wirklich’

‘Einfach’

- Focus particle (‘only’, ‘simply’)
- Analysis based on exclusion of alternatives (Beltrama 2021)

‘Echt/wirklich’

- VERUM focus, ‘really’ (Repp 2013, Romero 2015)
- Focus on truth conditions
- Wide (sentential) scope -> early position in predicative phrase
- Intensifying function could be derived similar to ‘simply’?

Identifying Intensifiers

Intensifiers are an open word class

- Our list of intensifiers contains 174 not listed by Claudi (2006) – 46 of Claudi's were not found in our corpus
- Most frequent: 'so' (see also Schumann 2021)
- Rapid change from adjective, adverb, focus particle to intensifier (but also other word classes)
- Well defined typical position before a gradable adjective

→ machine learning classifier

Data

- Part of TwiBloCoP (Twitter+Blog Corpus – Parenting)
- Automatic markup of all frequent intensifiers
- Manual annotation correction to include all intensifiers

```
#Text=Immer wieder , mal unbewusst , mal ganz deutlich .
45-1    1986-1991   Immer   _
45-2    1992-1998   wieder  _
45-3    1999-2000   ,   _
45-4    2001-2004   mal _
45-5    2005-2014   unbewusst   _
45-6    2015-2016   ,   _
45-7    2017-2020   mal _
45-8    2021-2025   ganz    Intensifier
45-9    2026-2034   deutlich    _
45-10   2035-2036   .   _
```

Statistical machine learning

- BIO-annotation for intensifying phrases (*so was von*)
- POS-tagging
- 80% training / 20% testing
- CRF classifier
  - Text, POS, casing (lower/upper) for current, previous and next token
- Baseline 1: SoMeWeTa POS tagger; PTKIFG (intensifying particle) tag
- Baseline 2: known intensifier before an adjective

Statistical ML: Results

	Macroaverage Precision	Macroaverage Recall	Macroaverage F_1
Baseline 1	0.45	0.49	0.47
Baseline 2	0.41	0.51	0.44
CRF classifier	**0.84**	**0.77**	**0.80**

Intensifier Classification: BERT

- Pretrained BERT-base from Huggingface
- Trained on 20k examples (as a prototype) split into 70% Train, 15% test and 15% validation datasets (Twitter only)

Epoch	Training Loss	Validation Loss	Overall Precision	Overall Recall	Overall F1	Overall Accuracy
1	0.001800	0.003829	0.985060	0.988670	0.986862	0.999102
2	0.001400	0.001995	0.990353	0.992003	0.991177	0.999434
3	0.000400	0.002288	0.990685	0.992336	**0.991510**	0.999451

RUHR
UNIVERSITÄT
BOCHUM

RUB

Summary

Intensifiers

- Intensifiers are extremely frequent, variable, and constantly changing
- Intensifiers differ in their expressive value
- More informative intensifiers are less frequent and combine with fewer adjectives
- "Vanilla" intensifiers precede more unusual ones and contribute to their interpretability
- There is a close semantic link between intensifiers and other particles, e.g., focus particles (*really*, *just*) and modal particles (*wohl*, French *bien*)

Intensifiers: Further computational models

- Classifier for intensifiers (using statistical machine learning or deep learning) ✓
- Cluster text by linguistic features such as intensifiers and particles for register analysis ✓
- Which information values should be used, how do we represent "information uniformity"?
- Train deep learning models to predict intensifiers and their order
- Predict intensifiers' expressivity

How do we capture creativity?

völligst
herrlich
abnormalst
sohoho, sio, doo, sool, soh
fking, fkin, f'n, fuckin'
väry
uhr
elends
megas
zer
umfucking
vollgas
def
definitiv ?
überkrass
hamma
gottes
gantz
sauig

ungeahnt
immens
vllig
verhurt
allzu
monstermäßig
töfte
superoberhypergigamaximal
derbst
aeusserst
sensationellst
absoulut
vohll
rundum, rundherum
sack
grotten
#arg
scheiß'
eminent

exorbitant
zucker
granatös
dollig
cutens
wrkl
kack
sterbens
zuupa
gscheid
piss
brauchebeimzeitungleseneinelesebrille
ganzschön
stroh
fluffing
dodal
oberhammer
verschissen
endlaser

Thank you!

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

ViCom

SFB 1475 Mo Re Metaphors of Religion

Faktenforum

Thank you to the DFG and BMBF for funding and to my students and collaborators.