tweeDe – A Universal Dependencies treebank for German tweets

Ines Rehbein, Josef Ruppenhofer and Bich-Ngoc Do

LiMo and IDS Mannheim

TLT 2019

INSTITUT FÜR
DEUTSCHE SPRACHE

Overview

- First German treebank for Twitter microtext
  - within the framework of Universal Dependencies
  - over 12,000 tokens from over 500 tweets

Outline of the talk

- Data selection and preprocessing
- Annotation process and IAA
- Baseline parsing results

Motivation

- Why another treebank for German?
- Treebanks for new language varieties
  - focus on **private communication**
  - highly informal, not carefully edited $\rightarrow$ spelling errors and ungrammatical structures
  - often lacks punctuation $\rightarrow$ problems for sentence segmentation
  - creative use of language $\rightarrow$ high ratio of OOV

Data Selection

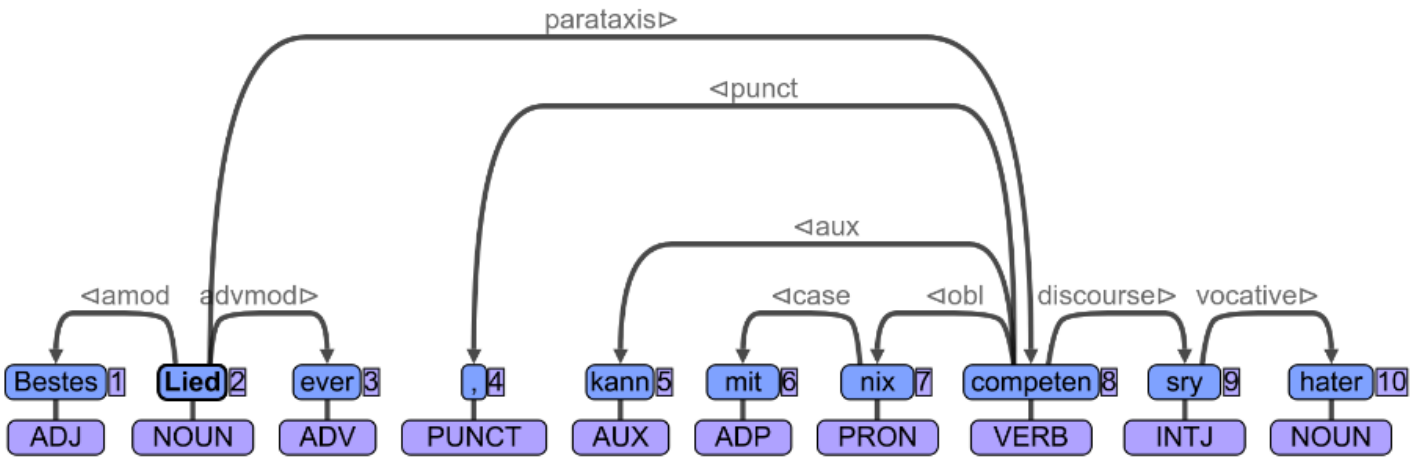
- Annotation of user-generated microtext is challenging
  - brevity of the messages
  - missing context information $\rightarrow$ highly ambiguous texts
- Solution: extract short communication threads
  - threads range in length from 2 – 34 tweets
  - annotators can see context, helps to resolve ambiguities

Data collection

- Data selection process:
  - Use German stop words as query terms to avoid topic bias
  - Retrieve conversation threads and download tweets
  - Only keep private communication between two or more users (manually remove ads, automatically generated messages, ...)
  - Treebank preserves the temporal order of the tweets in the same thread

Data structure and meta information

```
# tweet_id="969249808396537858" date="Sun Apr 14 09:57:57 +0000 2013" author="JD"
# text = "Trinkspiel" 10/10 mit @JzudemD Bestes Lied ever, kann mit nix competen sry hater
# sent_id = 15 tweet_sent_id = 2
# text = Bestes Lied ever, kann mit nix competen sry hater
1       Bestes  gut     ADJ     ADJA    Case=Nom|Degree=Sup|Gender=Masc|Number=Sing     2       amod    _       _
2       Lied    Lied    NOUN    NN      Case=Nom|Gender=Masc|Number=Sing        0       root    _       _
3       ever    ever    ADV     ADV     _       2       advmod  _       SpaceAfter=No
4       ,       ,       PUNCT   $,      _       8       punct   _       _
5       kann    können  AUX     VMFIN   Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   8       aux     _
_
6       mit     mit     ADP     APPR    _       7       case    _       _
7       nix     nix     PRON    PIS     Number=Sing|PronType=Neg        8       obl     _       _
8       competen        compete VERB    VVINF   VerbForm=Inf    2       parataxis       _       _
9       sry     sry     INTJ    ITJ     _       8       discourse       _       _
10      hater   hater   NOUN    NN      Case=Nom|Gender=Masc|Number=Plur        9       vocative        _       _
```



UD Annotatrix (Tyers et al. 2018), `https://github.com/jonorthwash/ud-annotatrix`

Segmentation

- Tweets now have up to 280 characters

Fach-Ing. R. Bernd @berndfachinsson · 30. Juli 2018

Antwort an @surfguard @Mathias59351078 @ArioMirzaie

Über einige amüsiere ich **mich** köstlich, bei manchen denke ich "hm" und bei wieder anderen bin ich entsetzt. Mit keinem einzigen hab ich irgendwas zu tun. **Wenn du mich wegen meiner Hautfarbe den Schuldigen zuordnest**, bist **du** ein Rassist.

1

Translation:

@surfguard @Mathias59351078 @ArioMirzaie Some make me laugh, some make me think ”hm“ and still others make me feel appalled. I dont have anything to do with any of them. If you blame me for the color of my skin, you’re a racist.

Segmentation

Raw tweet:

@surfguard @Mathias59351078 @ArioMirzaie Some make me laugh, some make me think "hm" and still others make me feel appalled. I dont have anything to do with any of them. If you blame me for the color of my skin, you're a racist.

Segmented text:

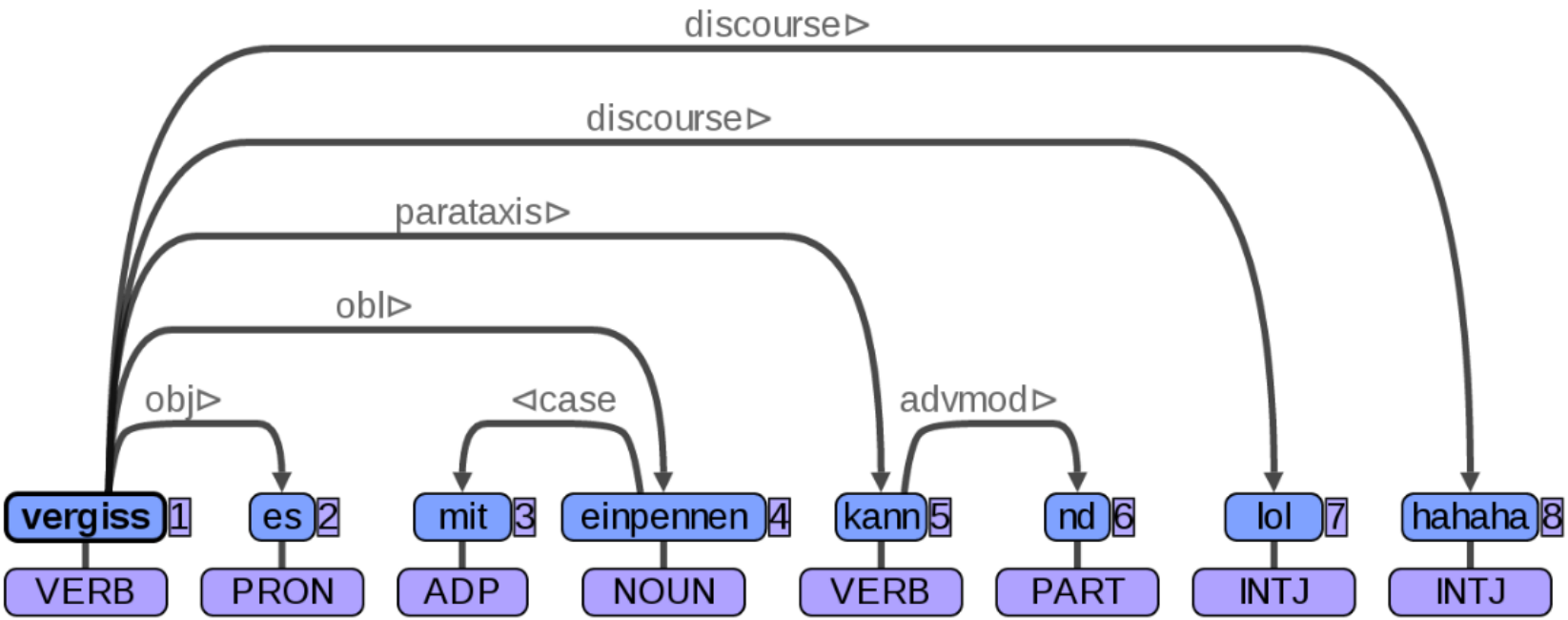
1 @surfguard @Mathias59351078 @ArioMirzaie

2 Some make me laugh, some make me think "hm" and still others make me feel appalled.

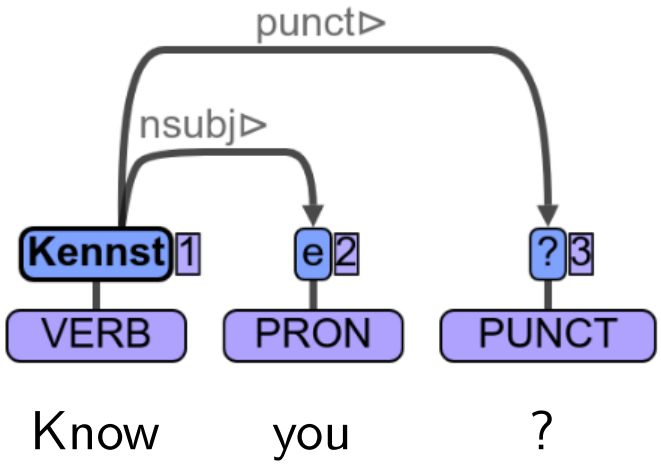
3 I dont have anything to do with any of them.

4 If you blame me for the color of my skin, you're a racist.

Segmentation



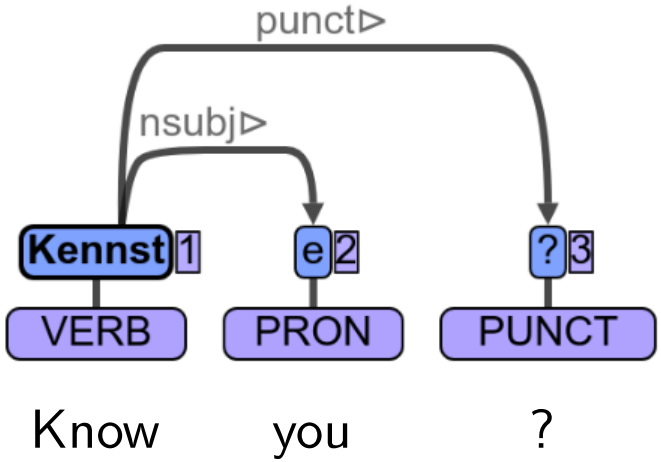
Tokenisation



(1) Kennste ? (raw)
Kennst e ? (tokenised)
kennen du ? (lemmatised)

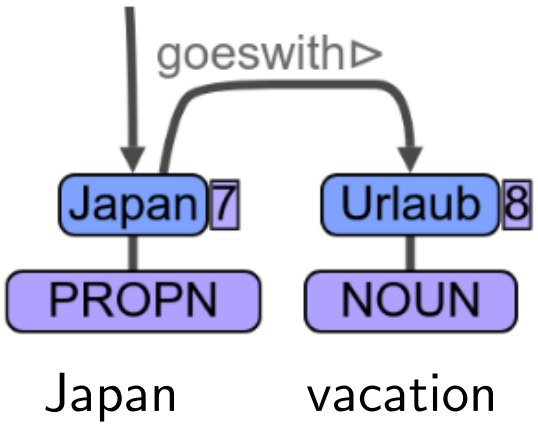
"Do you know that?"

Tokenisation



(2) Kennste ? (raw)
Kennst e ? (tokenised)
kennen du ? (lemmatised)

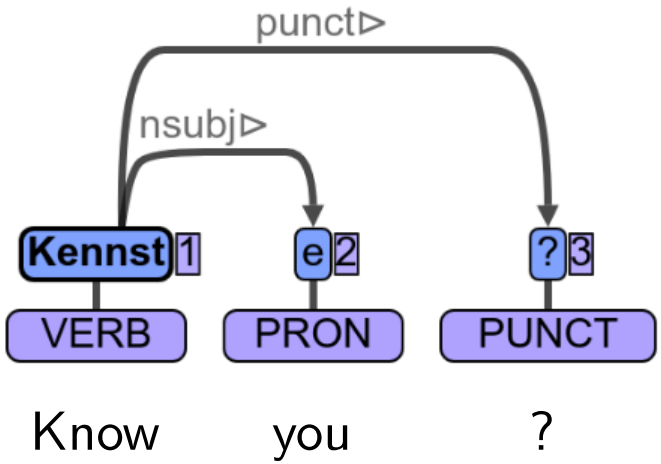
"Do you know that?"



(3) Japan Urlaub (raw)
Japan Urlaub (tokenised)
Japan Urlaub (lemmatised)

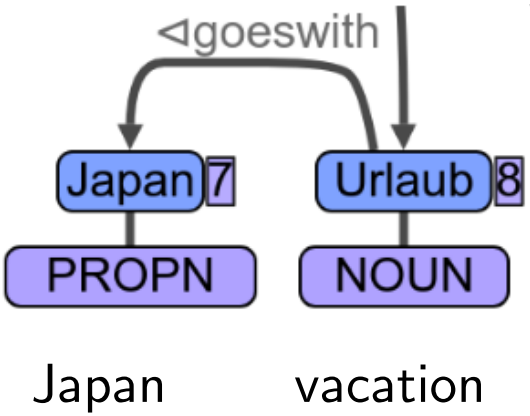
"vacation in Japan"

Tokenisation



(4) Kennste ? (raw)
Kennst e ? (tokenised)
kennen du ? (lemmatised)

"Do you know that?"



(5) Japan Urlaub (raw)
Japan Urlaub (tokenised)
Japan Urlaub (lemmatised)

"vacation in Japan"

Annotation

Data preannotated with UDPipe (Straka and Straková, 2017)

- Morphology and Parts-of-Speech
  - one annotator only
  - STTS (Schiller et al. 1995)
  - UD PoS (Petrov et al. 2012)
- Dependencies
  - all trees independently annotated by two annotators
  - disagreements resolved in discussion
  - consistency checks with scripts and DECCA (Dickinson and Meurers, 2003)

Annotation

Data preannotated with UDPipe (Straka and Straková, 2017)

- Morphology and Parts-of-Speech
  - one annotator only
  - STTS (Schiller et al. 1995)
  - UD PoS (Petrov et al. 2012)

- Dependencies
  - all trees independently annotated by two annotators
  - disagreements resolved in discussion
  - consistency checks with scripts and DECCA (Dickinson and Meurers, 2003)

- Inter-annotator agreement: 0.83 κ for labelled attachments, 0.89 κ for unlabelled attachments

Corpus statistics for 4 German UD treebanks

	text type	trees	tokens
UD-HDT	computer magazin	206,794	3,800,000
UD-TüBa	newspaper	104,787	1,959,474
UD-GSD	mostly web reviews	15,590	287,740
tweeDe	private tweets	1,301	12,073

Corpus statistics for 4 German UD treebanks

	text type	trees	tokens
UD-HDT	computer magazin	206,794	3,800,000
UD-TüBa	newspaper	104,787	1,959,474
UD-GSD	mostly web reviews	15,590	287,740
tweeDe	private tweets	1,301	12,073

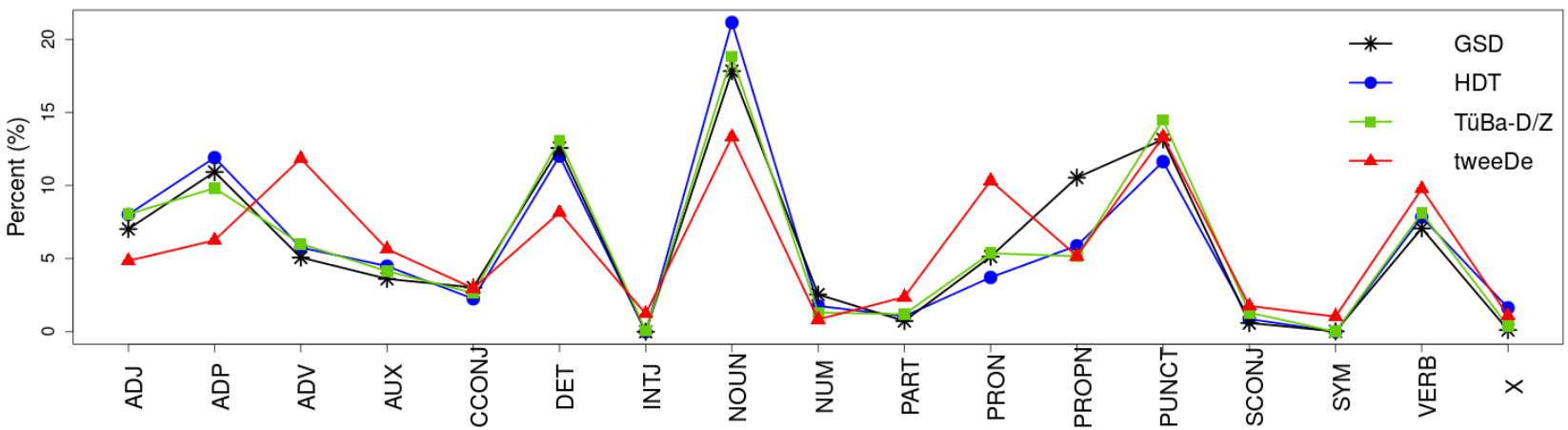


Figure: Distribution of UD PoS tags in four German UD treebanks.

tweeDe – Corpus statistics

tweeDe	# tweets	# tok	# vocab	OOV	lower
train	250	5,747	2,035	0	0
dev	69	1,917	861	520	479
test	200	4,409	1,661	1,157	1,034
total	**519**	**12,073**	**3,639**		

- **OOV**: number of out-of-vocabulary words wrt training set
- **lower**: OOV for lower-cased word forms.

Around 10% of the tweets include a non-projective tree structure.

Parsing baselines

- Parsing results for the Dozat parser (Dozat et al. 2017) on tweeDe, using gold POS

	PoS tagset	**dev** UAS	**dev** LAS	**test** UAS	**test** LAS
gold	UD	82.1	74.3	80.6	72.7
	STTS	73.5	63.0	70.3	60.8
	BOTH	**82.5**	**74.9**	**81.5**	**74.3**

Parsing baselines

- Parsing results for the Dozat parser (Dozat et al. 2017) on tweeDe, using gold POS

	PoS tagset	dev		test	
		UAS	LAS	UAS	LAS
gold	UD	82.1	74.3	80.6	72.7
	STTS	73.5	63.0	70.3	60.8
	BOTH	**82.5**	**74.9**	**81.5**	**74.3**

- Parsing results for the Dozat parser (Dozat et al. 2017) on tweeDe, using automatically predicted POS

	PoS tagset	dev		test	
		UAS	LAS	UAS	LAS
auto	UD	78.9	69.9	76.0	67.1
	STTS	72.9	63.2	71.2	62.6
	BOTH	**79.1**	**70.7**	**76.6**	**68.1**

Parsing baselines II

- Effect of more training data: Dozat parser trained on tweeDe and UD-GSD

	PoS tagset	**dev**		**test**	
		UAS	LAS	UAS	LAS
gold	UD	88.2	81.7	86.4	80.5
	STTS	85.2	77.3	81.4	74.0
	BOTH	**88.9**	**82.7**	**87.1**	**81.0**

Parsing baselines II

- Effect of more training data: Dozat parser trained on tweeDe and UD-GSD

	PoS tagset	**dev** UAS	**dev** LAS	**test** UAS	**test** LAS
gold	UD	88.2	81.7	86.4	80.5
	STTS	85.2	77.3	81.4	74.0
	BOTH	**88.9**	**82.7**	**87.1**	**81.0**

	PoS tagset	**dev** UAS	**dev** LAS	**test** UAS	**test** LAS
auto	UD	85.9	78.2	82.9	76.0
	STTS	84.9	76.4	82.3	74.8
	BOTH	**86.3**	**78.1**	**83.3**	**76.4**

Comparison to other Twitter dependency treebanks

	# token	# tweets	LAS	
EN (Foster et al. 2011)	n.a.	519⋆	67.3	♠
EN (Kong et al. 2014)	12,149	840	–	
EN-AAE (Blodgett et al. 2018)	3,072	250	56.1	♣
EN-MS (Blodgett et al. 2018)	3,524	250	67.7	♣
EN (Liu et al. 2018)	55,607	3,550	77.7	♣
IT (Sanguinetti et al. 2018)	124,410	6,712	81.5	♣
DE (this work)	12,073	519	68.1	♣
DE (this work)	+ UD-GSD		76.4	♣

⋆ Foster et al. only report # sentences, not # tweets

Seddah et al. (2012): phrase structure treebank

♠ Malt parser (Nivre et al. 2006)
♣ biaffine parser (Dozat & Manning 2017; Dozat et al. 2017)

Conclusions

- First German Twitter treebank, as a new training and testsuite for UD parsing
  - focus on informal communication
  - PoS, morphology, universal dependencies
- Parsing baselines
  - with the biaffine parser of Dozat et al. (2017): 83% (UAS) and 76% (LAS)
- Data will be available for research purposes: `http://www.cl.uni-heidelberg.de/~rehbein/resources/tweeDe.mhtml`

Thanks for listening!
Questions?