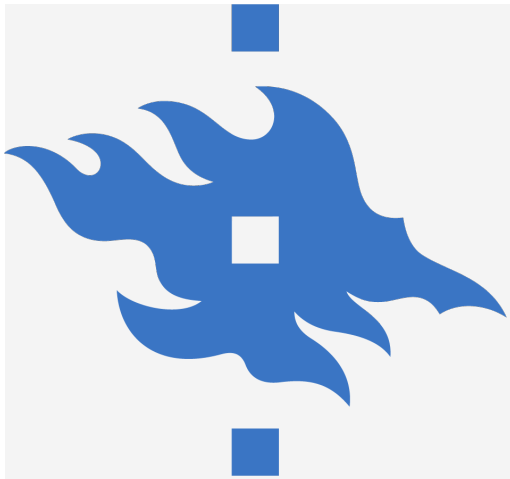
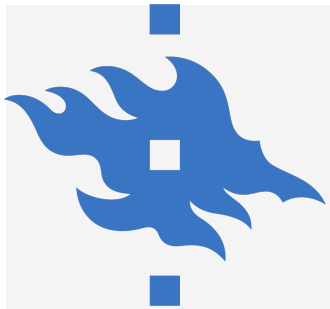


THE LOW SAXON LSDC DATASET
AT UNIVERSAL DEPENDENCIES

Janine Siewert and Jack Rueter



CONTENT

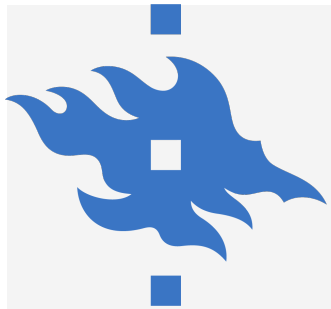
Introduction

Dataset content

Variation-related annotation challenges

Syntactic constructions

LANGUAGE AREA

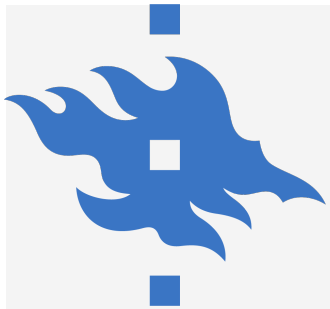


BACKGROUND INFORMATION

- Ca. 3–4 million speakers
- Official status in Germany, the Netherlands and parts of Brazil
- No interregional standard
- To some degree taught as a school subject in Germany
- Marginal use in media
- Specialisation at some universities, degree programme started in Oldenburg in autumn 2023

WRITING TRADITIONS

NWF:	Ziene olders hadden altied hard ewarkt en wazzen gezene leu in den naoberschop.
NNS:	Daor, kiek man ijs goud, ’t kan best wezen, dat ’t nog familie van die is.
DNS:	Arfest neem twe Kaarten to de eerst Klaß, un as ik daröver grote Ogen maak, lach he un meen, dat kunn darop staan, ik schull man instigen.
BRA:	Unn so wo de Doot dat den Fischer vertellt hett, isset ook ekåmen; dat ganze Dörp is uutstorven, man de Fischer is aarbliiwen unn issen riiken riiken Mann wåren, [...]
OFL:	Ik kann nich sä güt wiet lupen un dorumme schölle mik miene Fründin hier ne Parkbuchte friehulen.
DWF:	Eunige Dage später frogere de Magister, biu de veuer Johrestyien herren: Hiärmen sprank op, un de Magister mennte all, hai härr’ et wieten.



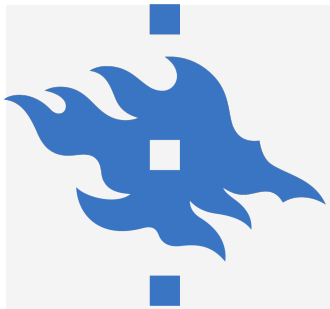
CONTENT

Introduction

Dataset content

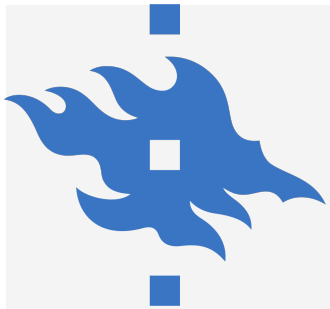
Variation-related annotation challenges

Syntactic constructions



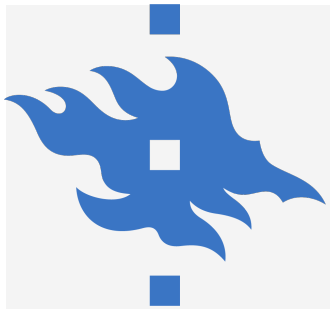
UD-DATASET

- Focus on the 19th and early 20th century
- Large overlap with my train, dev, and test data, but includes also Brandenburgish and Low Prussian
- Annotation of language change in progress



UD-DATASET CONTENT

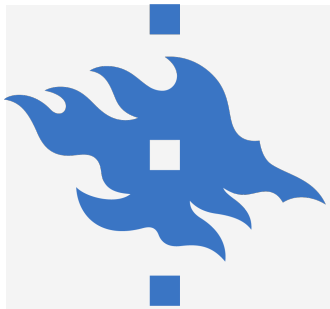
dialect	abbr	sent	token	lemma
Brandenburgish	BRA	48	1703	464
Dutch North Saxon	NNS	50	1,225	340
Dutch Westphalian	NWF	229	5,141	1,133
Eastphalian	OFL	50	1,575	460
German North Saxon	DNS	225	4,266	1,034
German Westphalian	DWF	238	4,471	1,012
Low Prussian	NPR	36	745	266
Mecklenburgish West–Pomeranian	MVP	124	3,505	833
total		1,000	22,631	5,542



SPELLING NORMALISATION AND LEMMATISATION

- Sentences are given both in original and in normalised spelling (*Nysassiske Skryvwyse* ‘New Saxon Spelling’)
- Both a Modern and a Middle Low Saxon lemma are provided

```
# sent_id = LSDC_0501_DNS_1911_HAM_hamborgsk_hein_godenwind_de_admirol_von_moskitonien
# text_orig = Hamborg, den twölften Dookmoond 1911.
# text = Hamborg, den twölvden doakmånd 1911.
1 Hamborg Hamborg PROPN _ Number=Sing 0 root _ lemma_gml=hamborch|SpaceAfter=No
2 , , PUNCT _ _ 5 punct _ _
3 den de DET _ Case=Acc|Definite=Def|Gender=Masc|Number=Sing|PronType=Art 5 det _
lemma_gml=dê,dê,dat
4 twölvden twelvde ADJ _ Case=Acc|Gender=Masc|Number=Sing|NumType=Ord 5 amod _ lemma_gml=twelfte
5 doakmånd doakmånd NOUN _ Case=Acc|Gender=Masc|Number=Sing 1 list _ lemma_gml=dākmânt
6 1911 1911 NUM _ NumType=Card 5 nummod _ SpaceAfter=No
7 .  .  PUNCT _ _ 1 punct _ _
```



CONTENT

Introduction

Dataset content

Variation-related annotation challenges

Syntactic constructions

VARIATION-RELATED ANNOTATION CHALLENGES -1

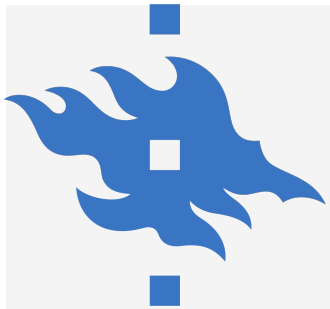
- Personal pronouns: 2nd person
  - *du/dû* ‘thou’
  - *jy/gî* ‘you’ – *jylüde/gîlûde* ‘you people’
  - (*see/sê*)
  - (*jim/gim*)

VARIATION-RELATED ANNOTATION CHALLENGES -2

- Grammatical gender:
  - Mostly three genders: feminine, masculine, neuter
  - Feminine and masculine gender have merged or are in the process of merging in several dialects
  - Variation in gender assignment
  - Lack of gender information in dictionaries

VARIATION-RELATED ANNOTATION CHALLENGES -3

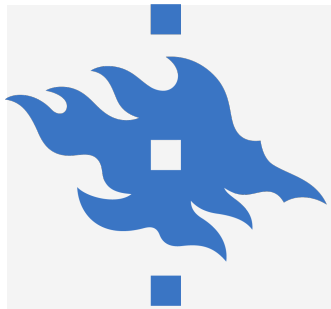
- Case inventory
  - **Nominative**, genitive, dative, **accusative**, (vocative?)
  - Ranging from 1/0 to 4
  - Remnants after certain prepositions:
    Mi weer de Sunn to grall bi ’n Läsen .
    me was the sun too bright at the.DAT.SG reading .
    ‘The sun was too bright for me while reading.’



VARIATION-RELATED ANNOTATION CHALLENGES -4

Mood inventory

- Indicative, imperative, subjunctive:
  et söl mi frögn, wank et bekäme
  it shall.PST.SBJV.3SG me please if-I it get.PST.SBJV-1SG
  ‘I would be happy if I got it.’
- Merger of indicative and subjunctive:
  du schusst man lewer to Huus gahn hebben
  you.SG shall.PST-2SG but rather to house go have
  ‘You had better gone home.’



CONTENT

Introduction

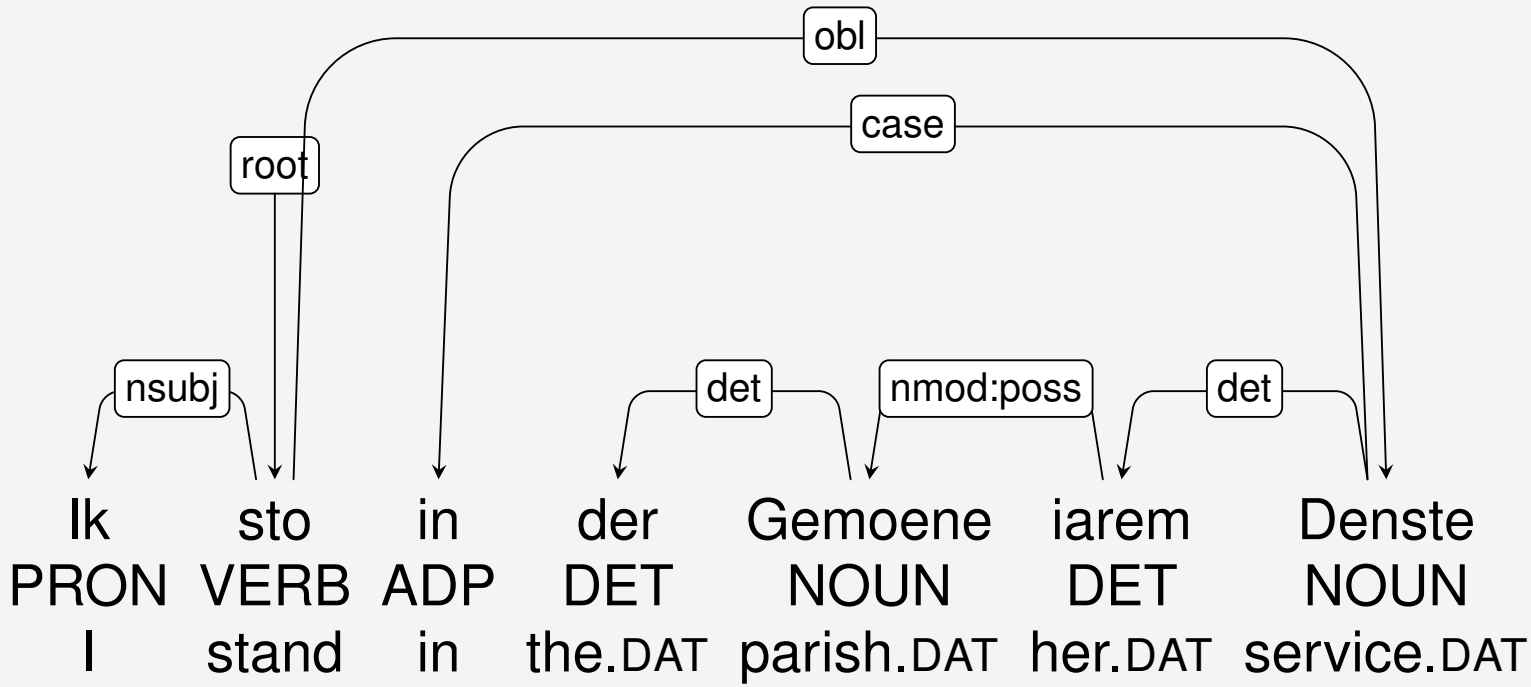
Dataset content

Variation-related annotation challenges

Syntactic constructions

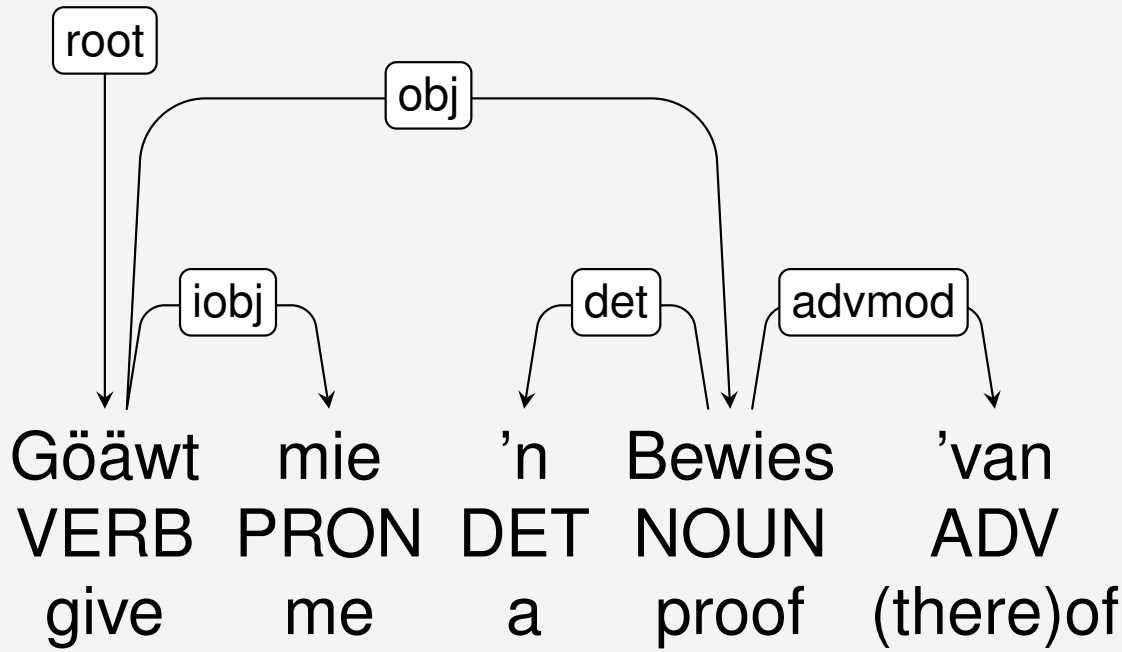
SYNTACTIC CONSTRUCTIONS

Possessive dative



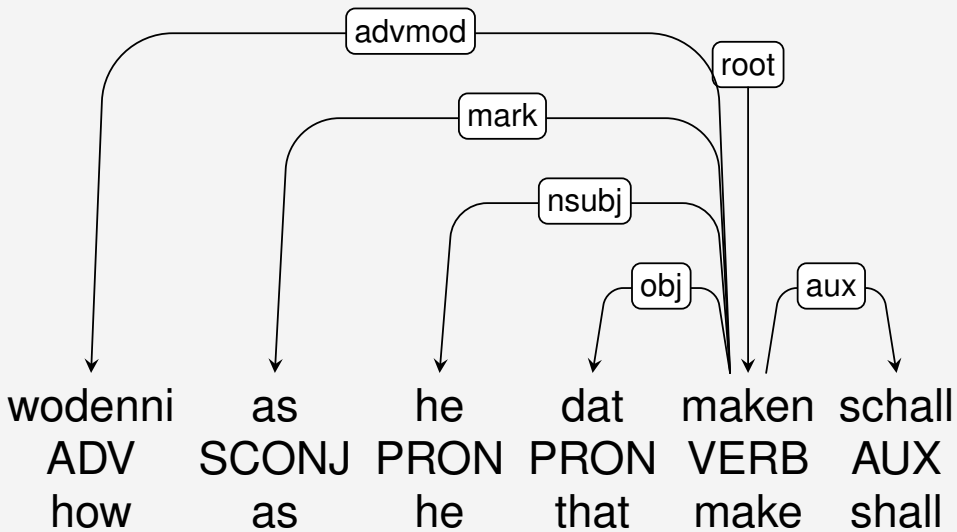
SYNTACTIC CONSTRUCTIONS

Pro-drop in separable adverbs



SYNTACTIC CONSTRUCTIONS

Complementiser doubling in subordinate interrogative clauses:
Un dạrmit secht de ol Mann em Beschêd, wodenni as he dat maken schall.
‘And with this, the old man tells him how he should do it.’



ACKNOWLEDGEMENTS

CorCoDial Project

“Corpus-based computational dialectology” – Academy of Finland project No. 342859

Dank jüm vöär’t luusteren!

`https://github.com/UniversalDependencies/UD_Low_Saxon-LSDC`