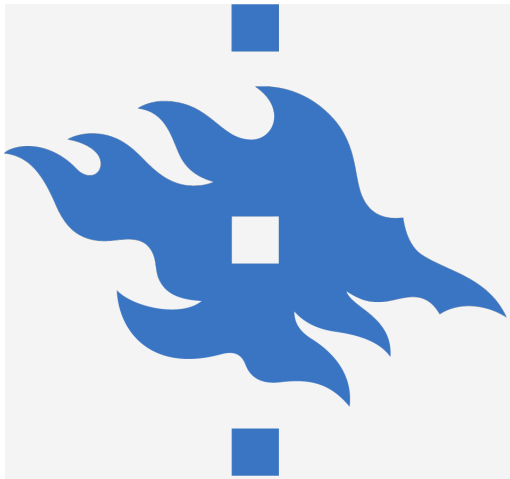
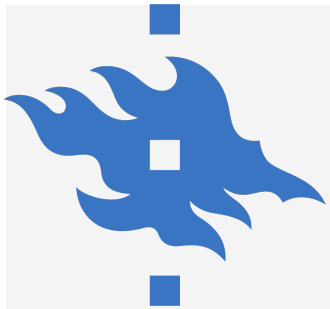


LOW SAXON
CORPUS-BASED DIALECTOMETRY

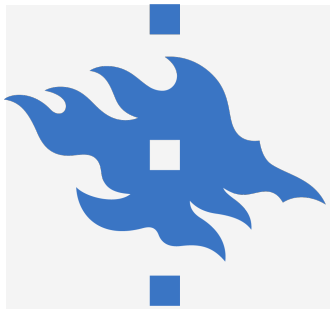
Janine Siewert

Supervisors: Yves Scherrer, Jörg Tiedemann, Martijn Wieling



CONTENT

LANGUAGE AREA

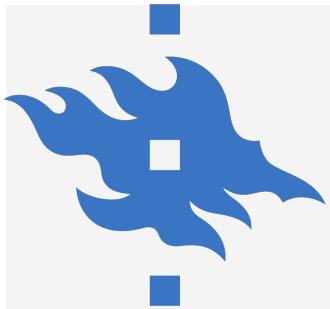


BACKGROUND INFORMATION

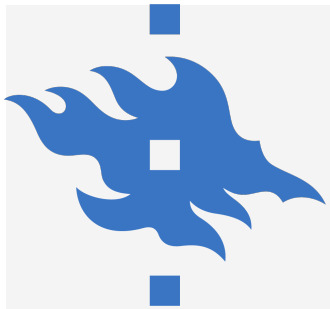
- Ca. 3–4 million speakers
- Official status in Germany, the Netherlands and parts of Brazil
- No interregional standard
- To some degree taught as a school subject in Germany
- Marginal use in media
- Specialisation at some universities, degree programme started in Oldenburg in autumn 2023

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.



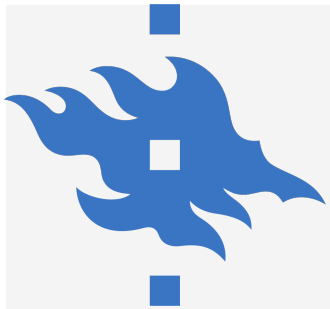
CONTENT



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

```
# 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 _ _
```



UD-DATASET CONTENT

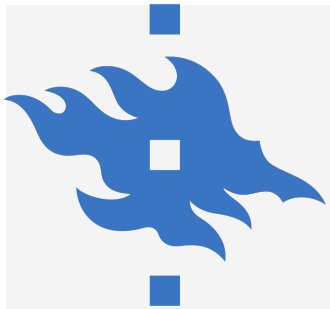
dialect	abbr	sent	token	lemma
Brandenburgish	BRA	48	1703	464
Dutch North Saxon	NNS	50	1,225	340
Dutch Westphalian	NWF	229	5,141	1,133
Eastphalian	OFL	50	1,575	460
German North Saxon	DNS	225	4,266	1,034
German Westphalian	DWF	238	4,471	1,012
Low Prussian	NPR	36	745	266
Mecklenburgish West–Pomeranian	MVP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

VARIATION-RELATED ANNOTATION CHALLENGES -1

- Personal pronouns: 2nd person
  - *du/dû* ‘thou’
  - *jy/gî* ‘you’ – *jylüde/gîlûde* ‘you people’
  - (*see/sê*)
  - (*jim/gim*)
- Grammatical gender:
  - Mostly three genders: feminine, masculine, neuter
  - Variation in gender assignment
  - Feminine and masculine gender have merged or are in the process of merging in several dialects

VARIATION-RELATED ANNOTATION CHALLENGES -2

- 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.’



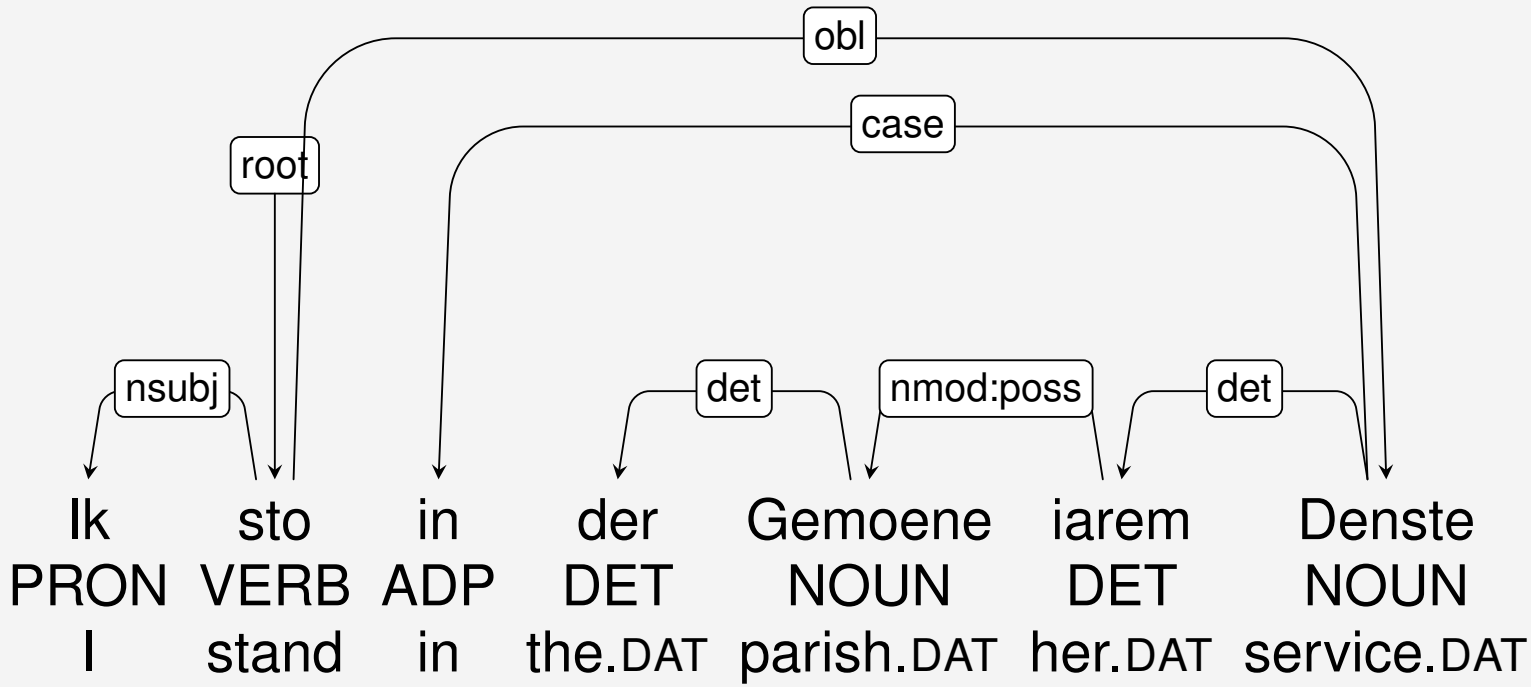
VARIATION-RELATED ANNOTATION CHALLENGES -3

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.’

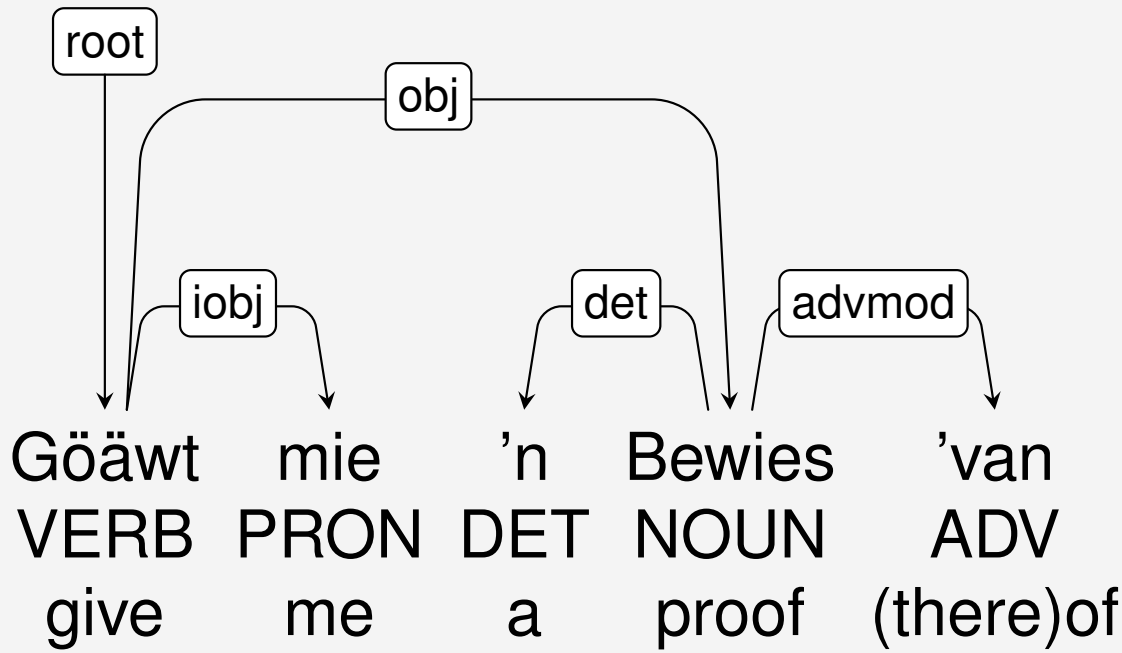
SYNTACTIC CONSTRUCTIONS

Possessive dative



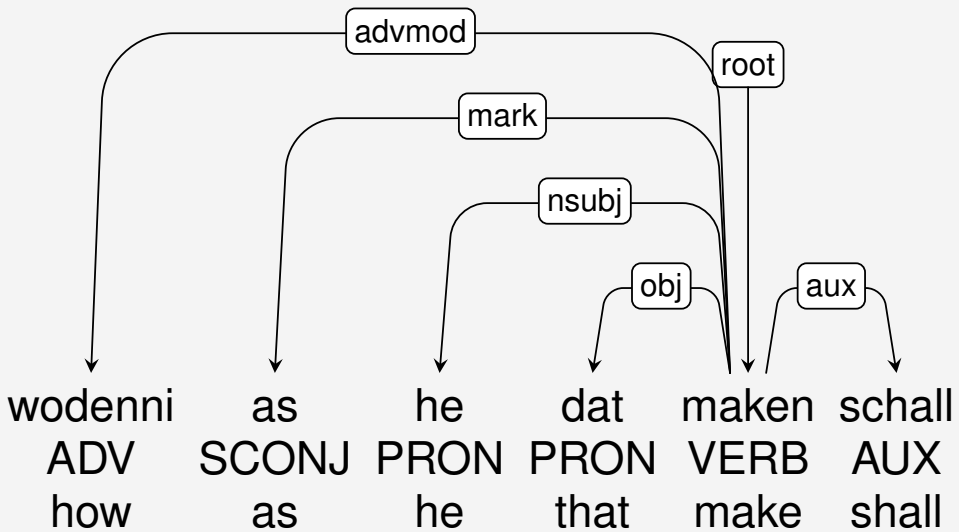
SYNTACTIC CONSTRUCTIONS

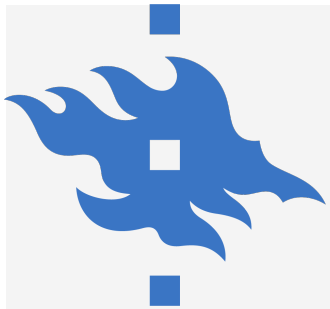
Pro-drop in separable adverbs



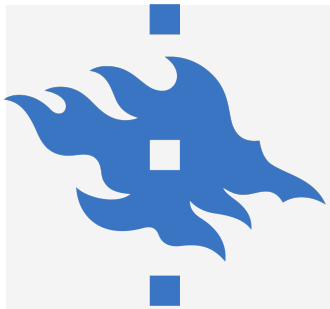
SYNTACTIC CONSTRUCTIONS

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.’





CONTENT



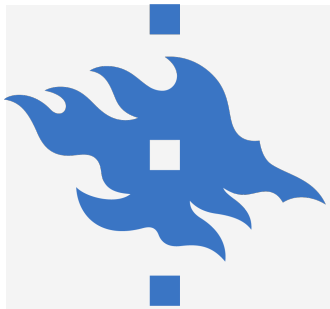
LEMMATISATION

Low Saxon is a West Germanic language spoken primarily in the north-eastern Netherlands and northern Germany (Moseley, 2010).

Occitan is a Gallo-Romance language spoken mainly in southern France and in parts of Italy and Spain (Bec, 1995).

Both languages are:

- **low-resourced** $\rightarrow$ limited amount of training data
- **non-standardized** $\rightarrow$ high levels of internal variation, both dialectal and orthographic

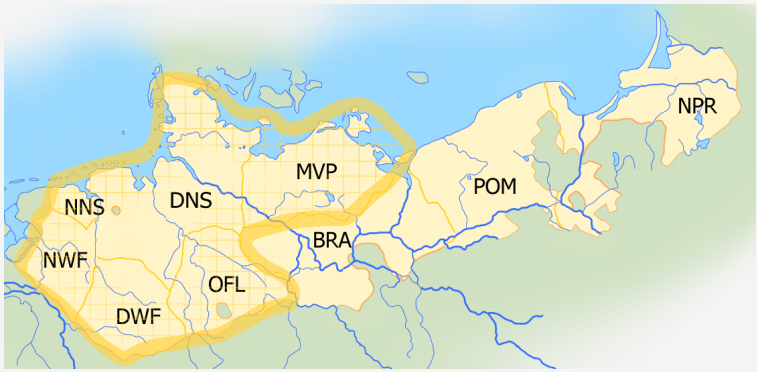


STRATEGIES

- **Training data amount**: is a large automatically annotated corpus more useful than a small gold standard corpus?
- **Training data specificity**: for a given dialect, is it better to use a general model trained on all dialects or a dialect-specific model?
- **Morphological information**: is PoS information better leveraged through sequential learning (e.g. Stanza (Qi et al., 2020)) or through joint learning (e.g. MaChAmp (van der Goot et al., 2021))?

LANGUAGES

Low Saxon



Mi weer de Sunn to grall bi 'n Läsen
ik weasen de sunne to gral by dat leasen
'The sun was too bright for me while reading.'

Occitan

Dins los bartasses bresilhavan un fum d' aucelons
dins lo bartàs bresilhar un fum d' aucelon
'In the bushes a multitude of birds were chirping.'

DATA AUGMENTATION

Low Saxon: leveraging a large corpus of historical Low Saxon (Peters, 2017) to train an initial lemmatization model
Occitan: ensembling a model based on a small gold corpus (Miletic et al., 2020) and a lexicon (Bras et al., 2020)

Dataset		Tokens*	Types*	T=L
Occitan	small	26.1	6.2	64.7
	large	2037.7	147.1	59.6
Low Saxon	small	19.3	6.0	39.7
	large	2431.9	166.6	39.1

**sizes in thousands. T=L: % of tokens identical to lemma.*

GENERAL RESULTS

Mean accuracy over three training runs, **all dialects**.

	Tool	Train	ALL	UNK	AMB
Low Sax. Occitan	MaChAmp	LARGE	91.8	68.5	92.2
		L+S	92.2	67.2	93.0
	Stanza	SMALL	**93.2**	**78.4**	**96.7**
		L+S	92.5	68.4	92.6
	MaChAmp	LARGE	**83.4**	30.2	85.2
		L+S	78.1	20.4	81.2
	Stanza	SMALL	80.5	**45.7**	**89.4**
		L+S	81.3	20.1	82.2

DIALECT-SPECIFIC RESULTS – OCCITAN

Mean accuracy over three training runs.

		Gascon			Lemosin			
Tool	Train	ALL	UNK	AMB	Train	ALL	UNK	AMB
MaChAmp	L+S	89.7	57.0	90.3	L+S	**90.9**	**74.4**	94.3
	L+GAS	88.9	54.4	89.6	L+LEM	87.6	64.3	92.7
Stanza	SMALL	**90.7**	**77.8**	**91.5**	SMALL	90.6	72.6	**99.2**
	L+S	90.1	67.5	89.6	L+S	89.8	66.7	92.7

		Lengadocian			Provençau			
Tool	Train	ALL	UNK	AMB	Train	ALL	UNK	AMB
MaChAmp	L+S	93.1	69.9	92.8	L+S	91.7	54.7	95.1
	L+LEN	92.6	68.3	92.3	L+PRO	86.6	52.0	89.5
Stanza	SMALL	**94.4**	**81.3**	**96.5**	SMALL	**92.8**	**74.9**	**98.5**
	L+S	93.7	71.5	93.0	L+S	92.1	54.7	93.5

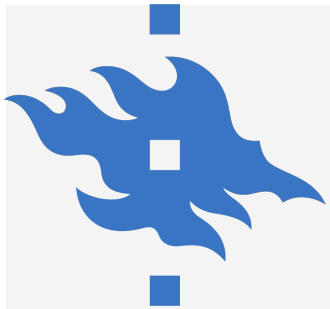
DIALECT-SPECIFIC RESULTS – LOW SAXON

Mean accuracy over three training runs.

		Dutch Low Saxon				German North Low Saxon		
Tool	Train	ALL	UNK	AMB	Train	ALL	UNK	AMB
MaChAmp	L+S	77.5	11.1	82.4	L+S	86.8	30.5	**90.3**
	L+DLS	76.3	10.6	81.2	L+NLS	82.6	**33.3**	85.3
Stanza	SMALL	**80.4**	**21.3**	**84.4**	SMALL	84.8	**33.3**	89.0
	L+S	78.9	14.3	82.0	L+S	**89.6**	30.5	89.0

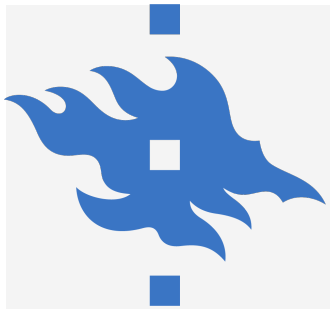
German South Low Saxon

Tool	Train	ALL	UNK	AMB
MaChAmp	L+S	74.0	45.4	74.5
	L+SLS	72.7	42.4	73.6
Stanza	SMALL	78.1	**47.0**	**79.4**
	L+S	**79.7**	33.3	78.4

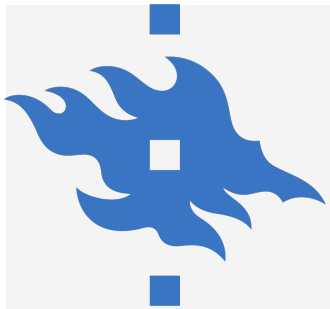


CONCLUSIONS

- Models trained on small, gold annotated corpora outperform models trained on larger amounts of automatically annotated silver corpora → **annotation reliability?**
- Models trained on all dialects outperform dialect-specific models → **amount of training data**?
- Using PoS in a sequential approach outperforms joint learning → **reliability of the PoS information**?
- Results on Low Saxon systematically lower than on Occitan → **higher degree of variation in the Low Saxon data**?



CONTENT



AUXILIARY AND MODAL VERBS

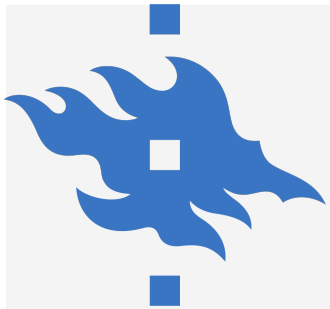
Two groups of auxiliary or modal verbs:

- **future auxiliaries** (*wērden* ‘to become’, *schō̈len* ‘shall’ and *willen* ‘will’)
- **models of permission, prohibition and obligation** (*dörven* ‘may, dare’, *dō̈ren* ‘dare’, *mo̊̂ten* ‘must’ and *mō̈gen* ‘may’) on the other.

For comparison, we also include the verbs *dôn* ‘to do’, *hebben* ‘to have’, *künnen* ‘can’, and *wēsen* ‘to be’.

DATA

Abbr.	Variety	Time span	Tokens
MLS	Middle Low Saxon	1200–1650	1 406 979
DLS1	Dutch Low Saxon	1800–1939	147 212
DLS2	Dutch Low Saxon	1980–2022	393 619
NLS1	German North LS	1800–1939	1 008 851
NLS2	German North LS	1980–2022	103 568
SLS1	German South LS	1800–1939	371 611
SLS2	German South LS	1980–2022	416 686



DATA ENCODING

- Three layers of annotation: Lemmas, PoS tags and dependency relations.
- The word vectors were trained on the whole dataset using fastText's (Bojanowski et al., 2016) skipgram model with a vector length of 100 and subwords following these two set-ups: lemma + dependency relation (e.g., *dörven_aux*), and lemma + PoS tags (e.g. *wērden_AUX*).
- Python library NumPy to measure the Euclidean distance between the resulting word vectors.

RESULTS – WĒRDEN, WITH DEPENDENCY RELATION

MLS	DLS1	DLS2	NLS1	NLS2	SLS1	SLS2
wēsen	wēsen	wēsen	wēsen	wēsen	mo̊̂ten	wēsen
hebben	dörven	dō̈ren	**schō̈len**	**schō̈len**	künnen	mo̊̂ten
künnen	dō̈ren	dörven	dörven	dörven	dörven	**schō̈len**
willen	mō̈gen	mo̊̂ten	mo̊̂ten	künnen	wēsen	mō̈gen
mo̊̂ten	mo̊̂ten	mō̈gen	künnen	dō̈ren	dō̈ren	dörven
dō̈ren	hebben	künnen	dō̈ren	hebben	**schō̈len**	hebben
dôn	künnen	dôn	hebben	mo̊̂ten	dôn	künnen
schō̈len	dôn	**schō̈len**	willen	willen	mō̈gen	dō̈ren
mō̈gen	**schō̈len**	willen	mō̈gen	dôn	willen	willen
dörven	willen	hebben	dôn	mō̈gen	hebben	dôn

RESULTS – WĒRDEN, WITH POS

MLS	DLS1	DLS2	NLS1	NLS2	SLS1	SLS2
wēsen	hebben	wēsen	wēsen	wēsen	dören	wēsen
dôn	dören	dören	**schölen**	künnen	wēsen	dörven
willen	mögen	dörven	hebben	dören	môten	hebben
hebben	wēsen	dôn	môten	hebben	künnen	mögen
mögen	dörven	hebben	dören	**schölen**	dörven	**schölen**
dören	môten	môten	künnen	dörven	**schölen**	willen
schölen	künnen	**schölen**	dörven	dôn	dôn	dôn
môten	dôn	künnen	willen	mögen	hebben	môten
künnen	willen	mögen	dôn	willen	mögen	künnen
dörven	**schölen**	willen	mögen	môten	willen	dören

RESULTS – DÖRVEN, WITH DEPENDENCY RELATIONS

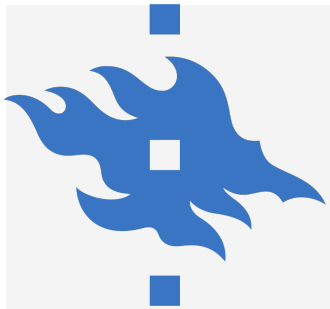
MLS	DLS1	DLS2	NLS1	NLS2	SLS1	SLS2
mô̲ten	mô̲ten	**dō̲̈ren**	künnen	schō̲̈len	schō̲̈len	mô̲ten
dō̲̈ren	**dō̲̈ren**	*mō̲̈gen*	mô̲ten	**dō̲̈ren**	mô̲ten	willen
willen	*mō̲̈gen*	willen	schō̲̈len	künnen	willen	künnen
mō̲̈gen	künnen	mô̲ten	willen	willen	**dō̲̈ren**	dôn
schō̲̈len	dôn	künnen	**dō̲̈ren**	mô̲ten	künnen	**dō̲̈ren**
künnen	wērden	schō̲̈len	hebben	hebben	wēsen	wēsen
hebben	schō̲̈len	dôn	wēsen	*mō̲̈gen*	dôn	hebben
dôn	hebben	wēsen	*mō̲̈gen*	dôn	hebben	schō̲̈len
wēsen	wēsen	hebben	dôn	wēsen	*mō̲̈gen*	wērden
wērden	willen	wērden	wērden	wērden	wērden	*mō̲̈gen*

RESULTS – DÖRVEN, WITH POS

MLS	DLS1	DLS2	NLS1	NLS2	SLS1	SLS2
dören	**dören**	**dören**	willen	**dören**	möten	möten
möten	wēsen	willen	hebben	schölen	**dören**	**dören**
mögen	wērden	schölen	möten	hebben	künnen	künnen
willen	*mögen*	dôn	**dören**	*mögen*	dôn	dôn
künnen	möten	*mögen*	schölen	willen	wērden	willen
schölen	künnen	hebben	künnen	künnen	schölen	wēsen
dôn	hebben	künnen	dôn	wēsen	wēsen	hebben
hebben	dôn	wēsen	wēsen	dôn	*mögen*	schölen
wērden	willen	möten	wērden	möten	willen	*mögen*
wēsen	schölen	wērden	*mögen*	wērden	hebben	wērden

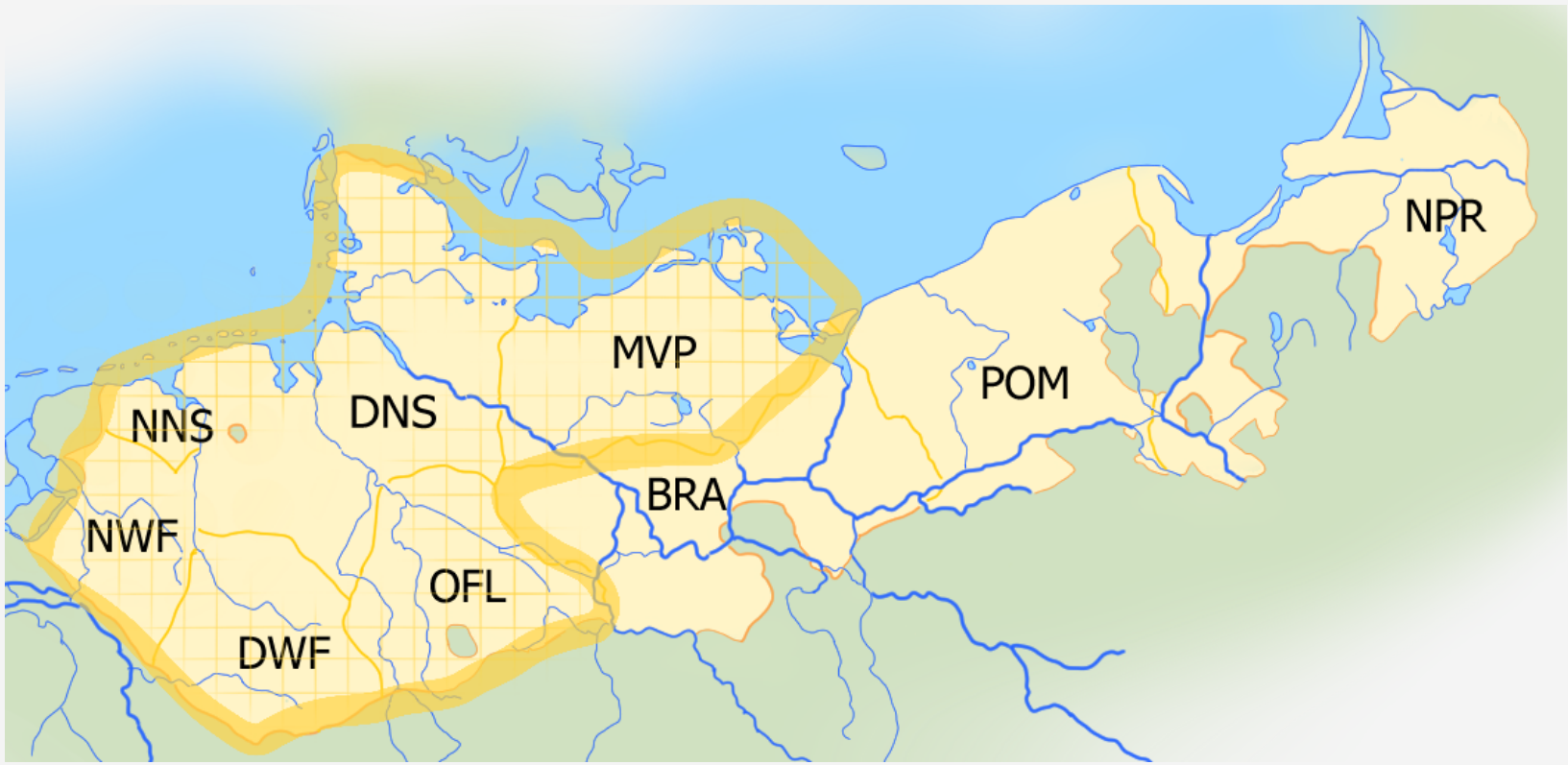
DISCUSSION

- The increased closeness of *schō̈len* to *wērden* in German Low Saxon is in line with the development of *wērden* into a future tense auxiliary.
- The decreased closeness of *willen*, at least for modern German Low Saxon, might show that the usage as a future auxiliary is in fact not very widespread.
- In the similarity of *dörven* to *dō̈ren* and *mō̈gen* we see a decrease in German Low Saxon. This might be related to the usage of Standard German *dürfen* and *mögen*.
- Moreover, a shift in the usage of negated *mô̂ten* from ‘must not / to not be allowed to’ to ‘do not need to’ as in German might explain the decreased similarity in NLS.

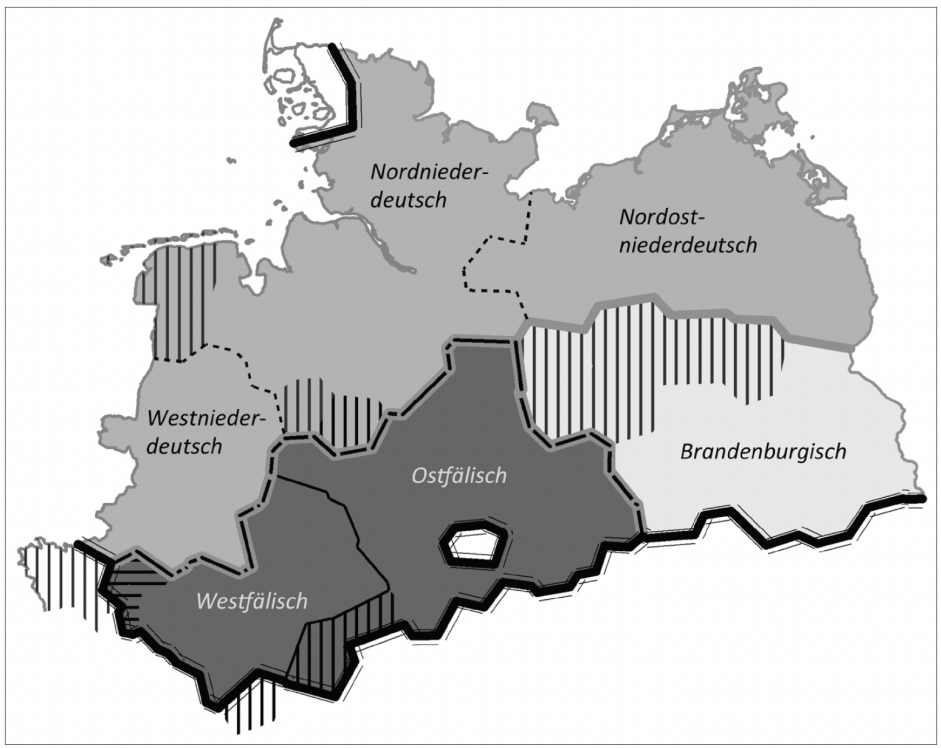


CONTENT

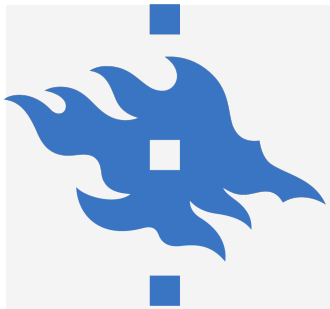
LOW SAXON DIALECTS INCLUDED



GERMAN LOW SAXON DIALECTS ACCORDING TO LAMELI



Lameli, Alfred (2016). Raumstrukturen im Niederdeutschen. Eine Re-Analyse der Wenkerdaten. *Jahrbuch des Vereins für niederdeutsche Sprachforschung*, 139, 131-152.



DIALECT DISTANCES

- Distances based on characters, PoS and morphological features
- What changes can be observed?
- Do the different levels produce different groupings?
- What role does the political border play?
- Do we find the traditional east-west division?

RESULTS – POS

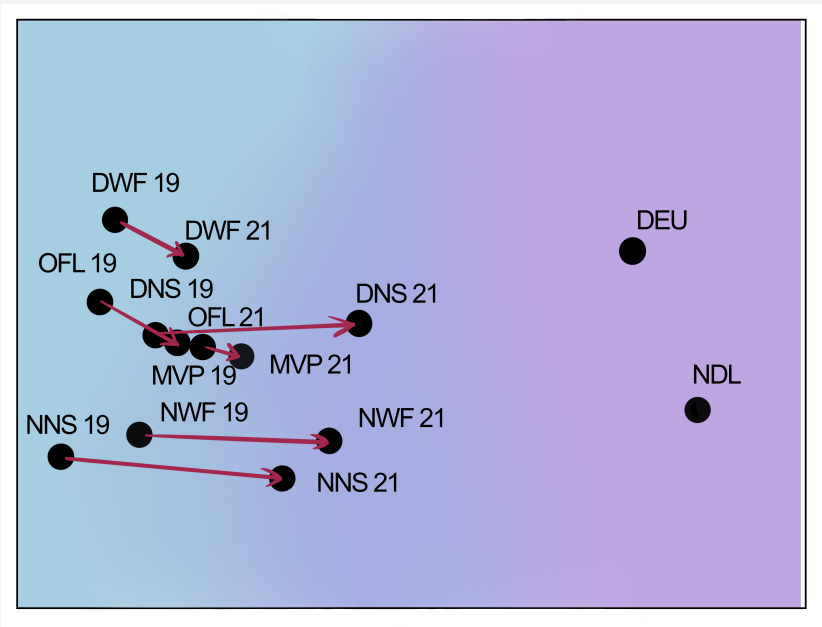


Figure: PCA

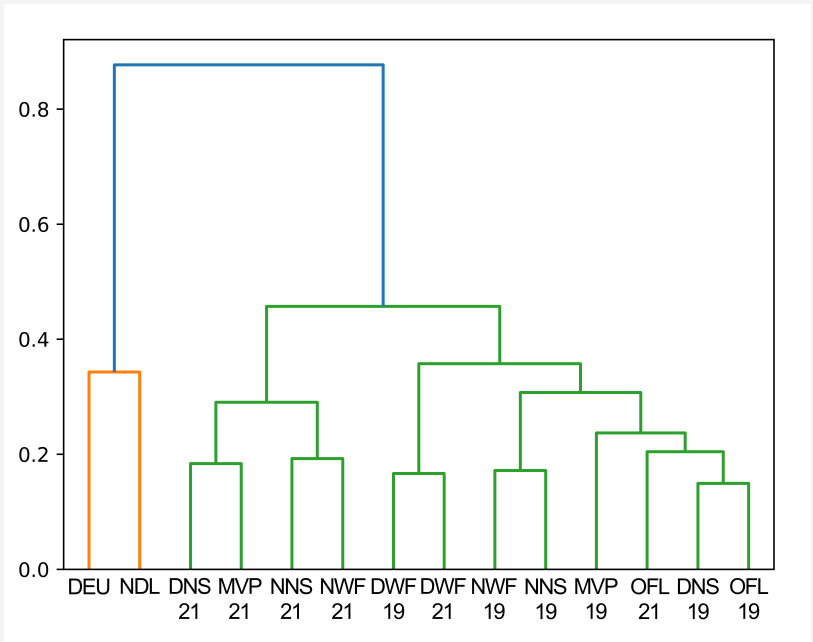


Figure: Hierarchical

RESULTS – POS + MORPH

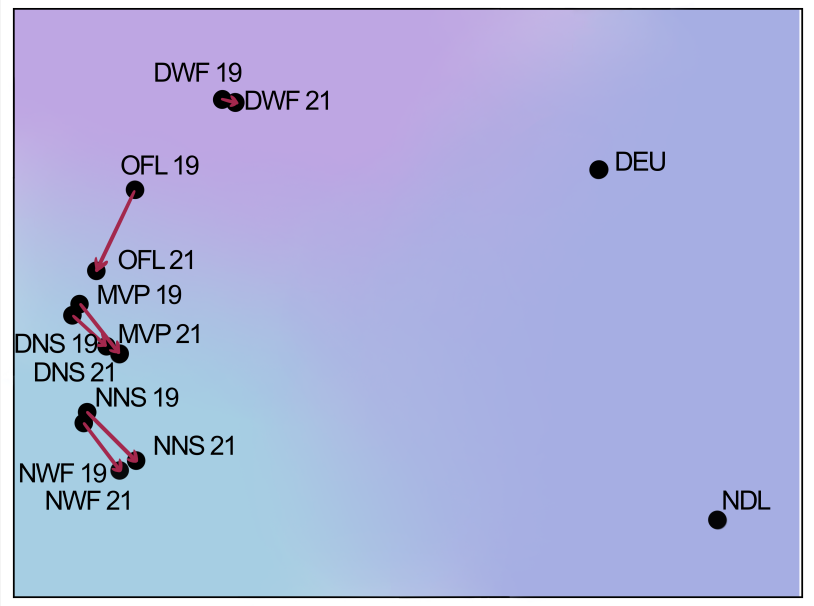


Figure: PCA

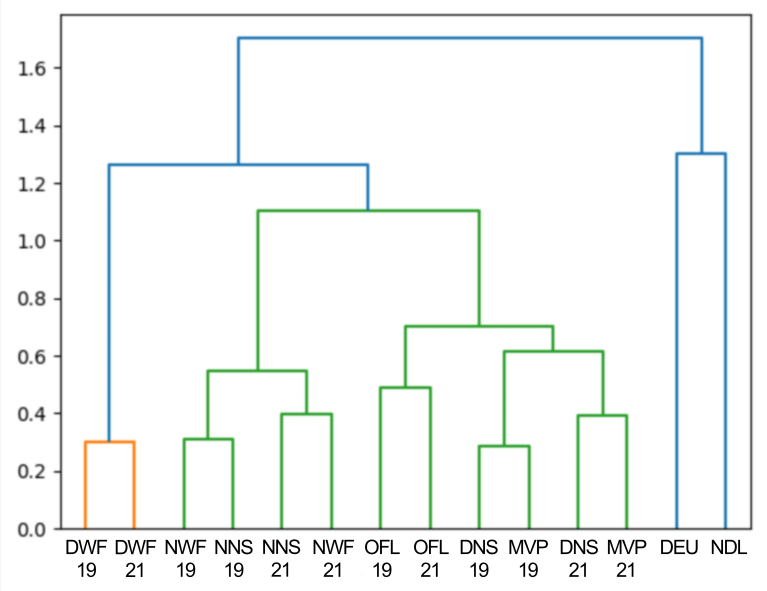


Figure: Hierarchical

RESULTS – POS + MORPH, GENDER=COM

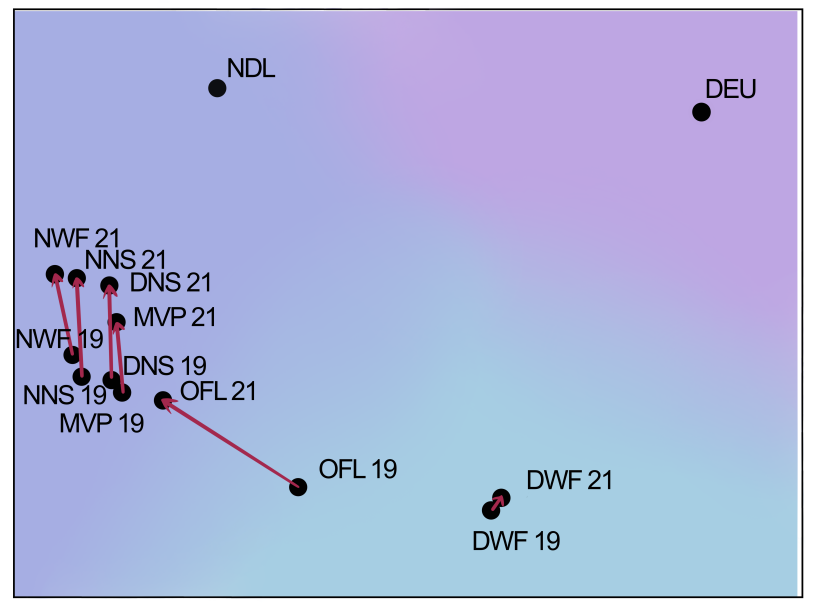


Figure: PCA

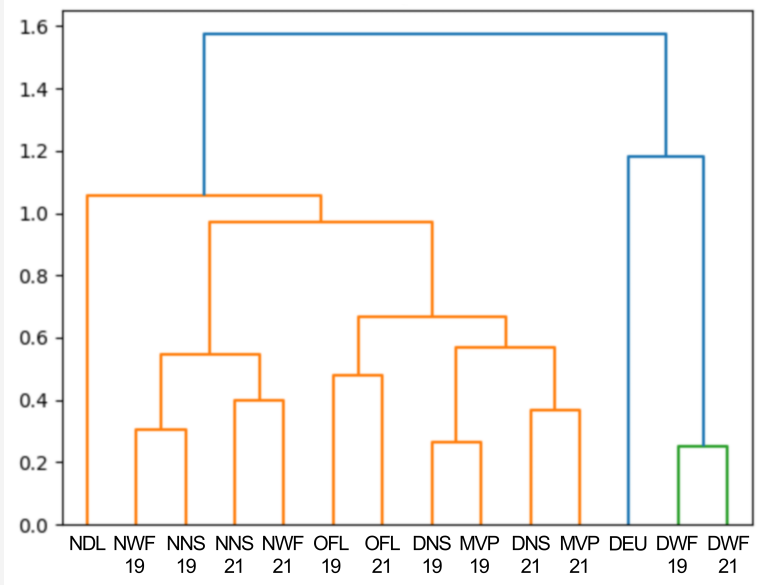
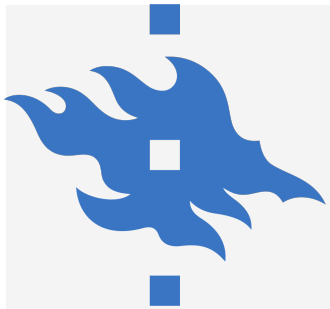


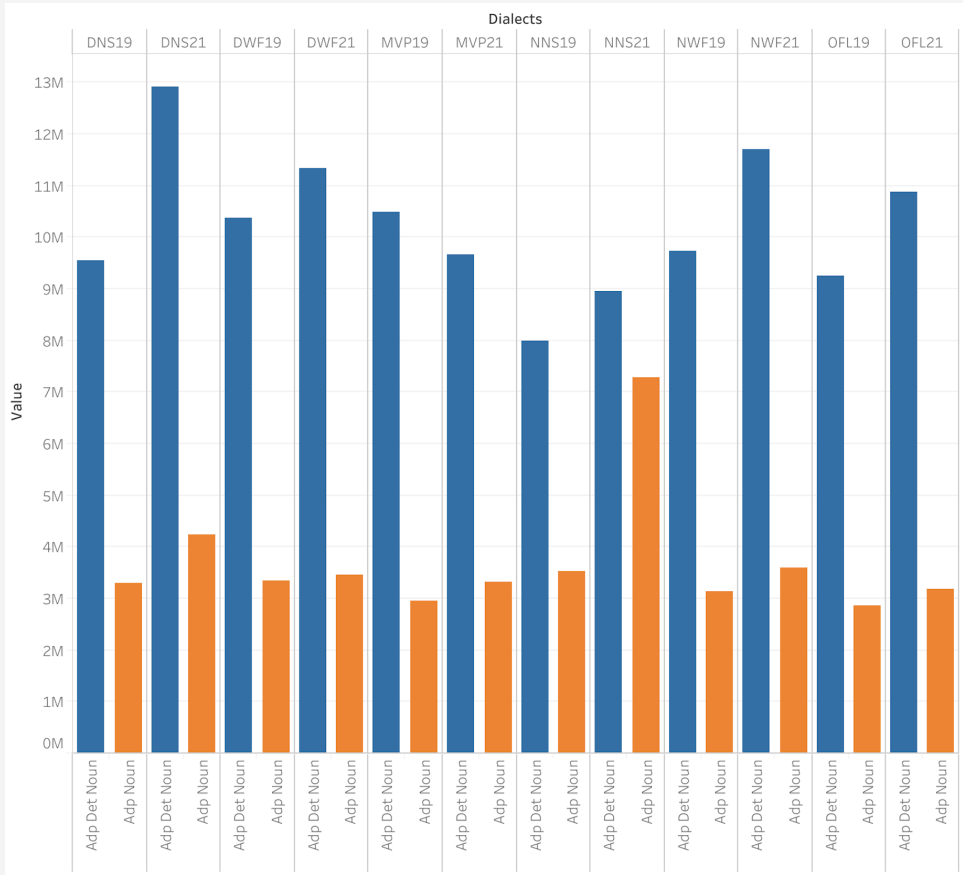
Figure: Hierarchical

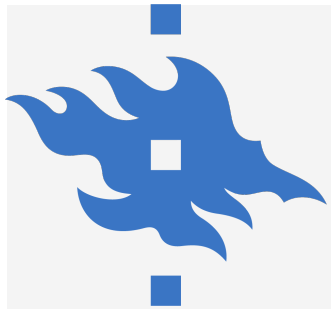


N-GRAM ANALYSIS

- ADP-DET bigram less frequent than in German and Dutch (see next slide)
- Case=Dat feature high values in German and German Westphalian
- IPP (Infinitivus Pro Participio) construction occurs in all dialects and is preferred in most of them
  No IPP: *Ik had dat doon kund.*
  With IPP: *Ik had dat doon künnen.*
  'I could have done that.'

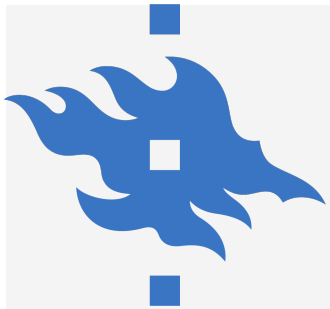
ADP-DET IN LOW SAXON





SUMMARY

- The political border plays a role but does not explain all developments
- The traditional east-west division does not become apparent
- Northern dialects from German increasingly resemble Dutch
- NL Westphalian closer to the northern dialects than to DE Westphalian
- NDS-NL clusters according to the period, not according to the dialect



DISCUSSION

- Limitations:
  - Different size of dialect regions
  - Data (and speaker) availability
- Current and future research:
  - Changing article usage in Low Saxon
  - Dialect distances based on dependency relations
  - Interpretable dialect classification, compare original and orthographically normalised data

ACKNOWLEDGEMENTS

CorCoDial Project

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

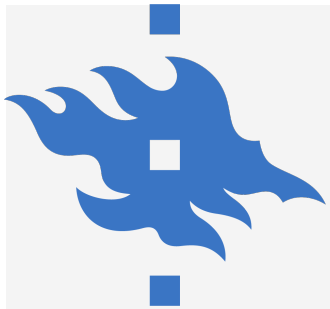
FoTran Project

“Found in Translation: Natural Language Understanding with Cross-lingual Grounding” – ECR funded project

UniDive COST Action

“Universality, diversity and idiosyncrasy in language technology” – COST Association, COST Action CA21167

Dank jüm vöär’t luusteren!

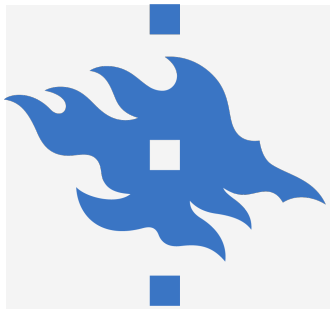


BIBLIOGRAPHY I

Pierre Bec. *La langue occitane*. PUF, 6th edition, 1995.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

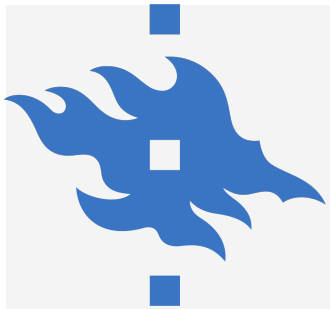
Myriam Bras, Marianne Vergez-Couret, Nabil Hathout, Jean Sibille, Aure Séguier, and Benazet Dazéas. Loflòc : Lexic obèrt flechit occitan. In Jean-François Courouau, editor, *Fidélités et dissidences (Actes du* XII*e congrès de l'Association Internationale d'Études Occitanes)*, pages 141–156, Albi, 2020. Centre d'Etude de la Littérature Occitane.



BIBLIOGRAPHY II

Aleksandra Miletic, Myriam Bras, Marianne Vergez-Couret, Louise Esher, Clamença Poujade, and Jean Sibille. A four-dialect treebank for Occitan: Building process and parsing experiments. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 140–149, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics (ICCL). URL `https://aclanthology.org/2020.vardial-1.13`.

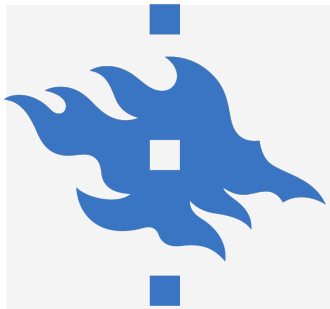
Christopher Moseley, editor. *Atlas of the World's Languages in Danger*. UNESCO Publishing, Paris, 3 edition, 2010. Online version: `http://www.unesco.org/culture/en/endangeredlanguages/atlas`.



BIBLIOGRAPHY III

Robert Peters. Das referenzkorpus mittelniederdeutsch/niederrheinisch (1200–1650). *Niederdeutsches Jahrbuch. Jahrbuch des Vereins für niederdeutsche Sprachforschung*, 140: 35–42, 2017.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020. URL `https://nlp.stanford.edu/pubs/qi2020stanza.pdf`.



BIBLIOGRAPHY IV

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-demos.22. URL https://aclanthology.org/2021.eacl-demos.22.