LOW SAXON INTERNAL VARIATION

AT THE ORTHOGRAPHIC, MORPHOLOGICAL AND SYNTACTIC LEVEL

Janine Siewert and Yves Scherrer

TABLE OF CONTENTS

INTRODUCTION: LOW SAXON DIALECT CONTINUUM

INTRODUCTION & BACKGROUND: WRITING SYSTEMS

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, unn siene Kinger leewen noch bett upp dissen Dach in Göttin unn sinn riike Lüüe.
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.

INTRODUCTION: LANGUAGE CONTACT

- Different majority languages: Dutch and German
- Noticeable influence on lexicon, phonetics (e.g. prosody)
- Some influence on morphology
  - Loss of old 2nd person singular
  - Preservation or loss of case inflection
- Influence on syntax still underresearched

TABLE OF CONTENTS

DATASET

- Low Saxon from two time periods (sentences):

	1800-1939	1980-2022
German North Saxon (DNS)	22,558	3,532
Mecklenburgish (MVP)	19,928	3,054
Dutch North Saxon (NNS)	1,805	16,950
German Westphalian (DWF)	16,945	15,128
Dutch Westphalian (NWF)	4,881	9,137
Eastphalian (OFL)	1,663	7,883

- Modern Dutch and German

LOW SAXON DIALECTS

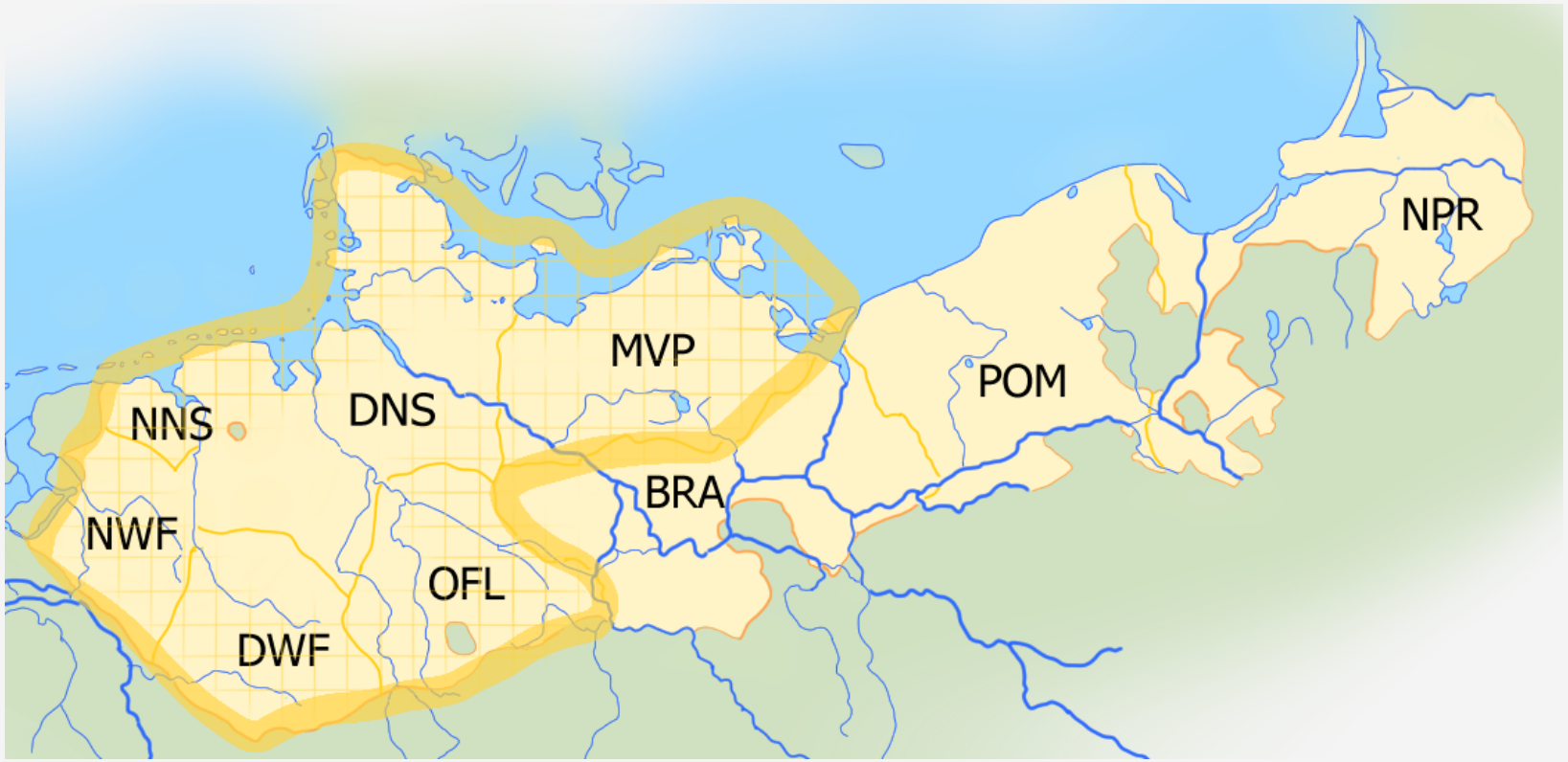


TABLE OF CONTENTS

DIALECT DISTANCES AND ANNOTATION

- Distances based on bigrams and trigrams of:
  - characters — orthography and phonology
  - PoS tags — syntax
  - PoS tags and morphological features — morphology

DIALECT DISTANCES AND ANNOTATION

- Distances based on bigrams and trigrams of:
  - characters — orthography and phonology
  - PoS tags — syntax
  - PoS tags and morphological features — morphology
- Semi-automatic annotation:
  - Stanza tagger trained on UD datasets in German, Dutch, Danish and Swedish in addition to manually corrected Low Saxon data
  - PoS accuracy 92% and feature accuracy 83% on held-out Low Saxon test sets
  - Also used for retagging German and Dutch

RESEARCH QUESTIONS

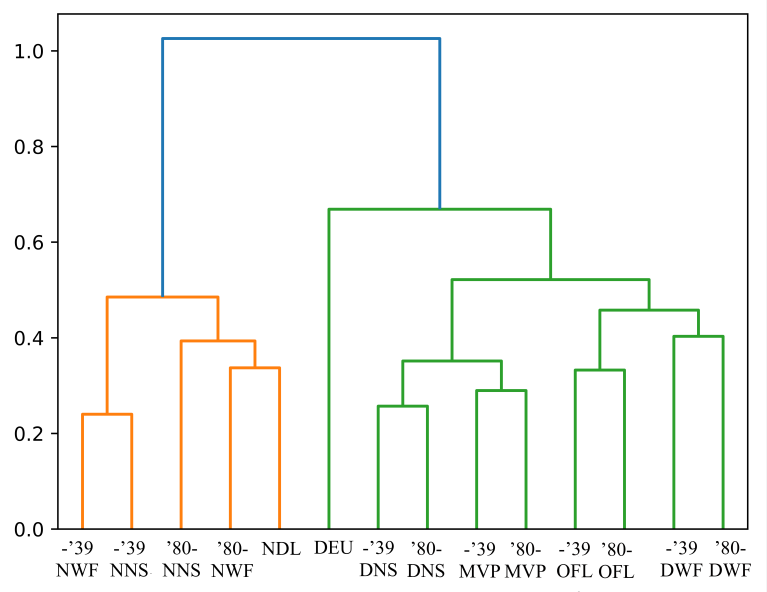
- Do the three approaches lead to different groupings?
- What changes can be observed from the 19th to the 21st century?
- Are the political border and the traditional east-west division visible in the data?

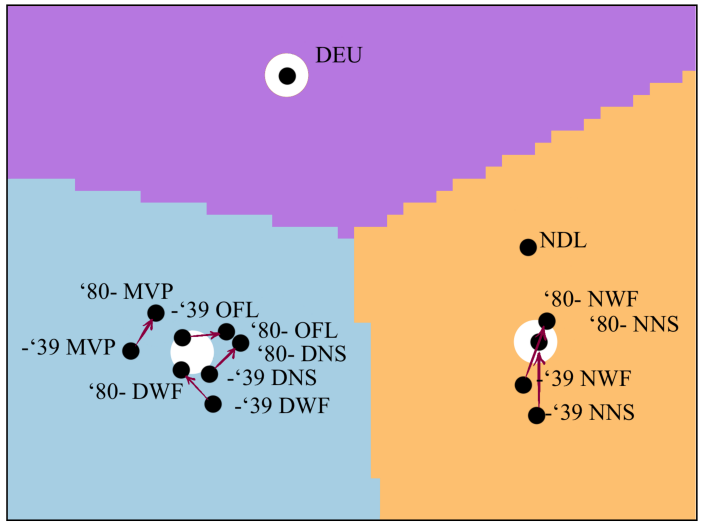
APPROACHES

- tf-idf normalised counts of 2-grams and 3-grams
  - exluding the following PoS tags:
    ’SYM’, ’_’, ’X’, consecutive ’PUNCT’
- K-means clustering
- Principal Component Analysis (PCA)
- Hierarchical clustering

TABLE OF CONTENTS

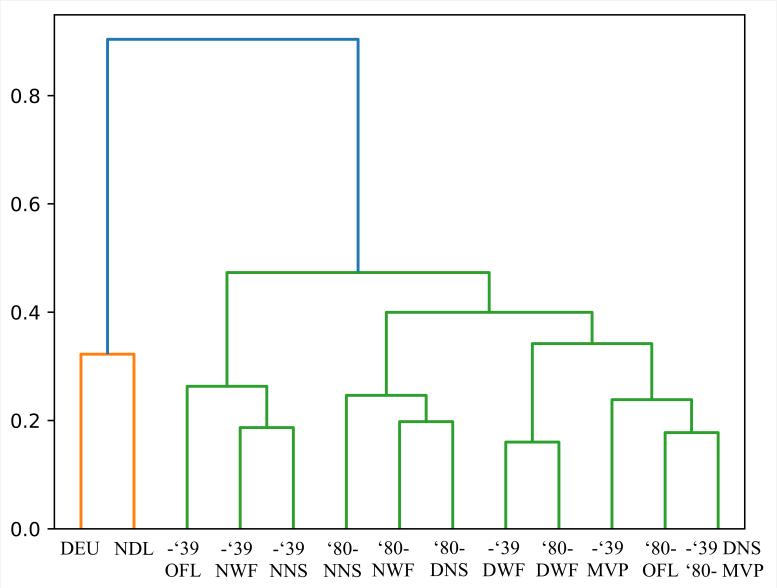
RESULTS – CHARACTER LEVEL

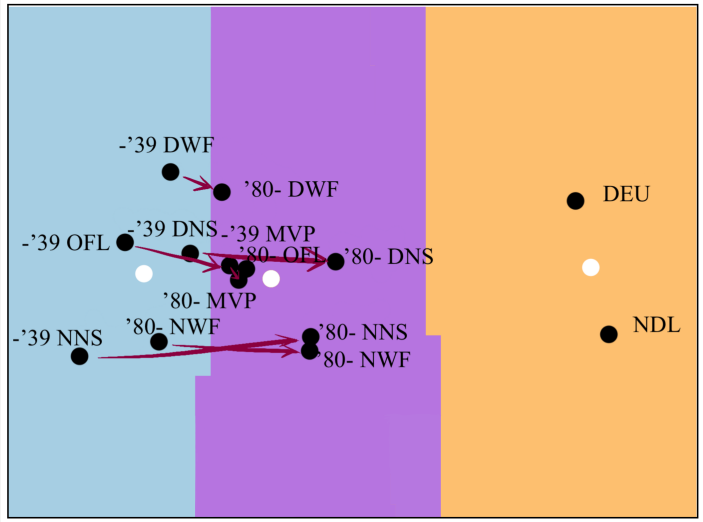




Dutch Low Saxon clusters with Dutch.
German Low Saxon remains separate from German.

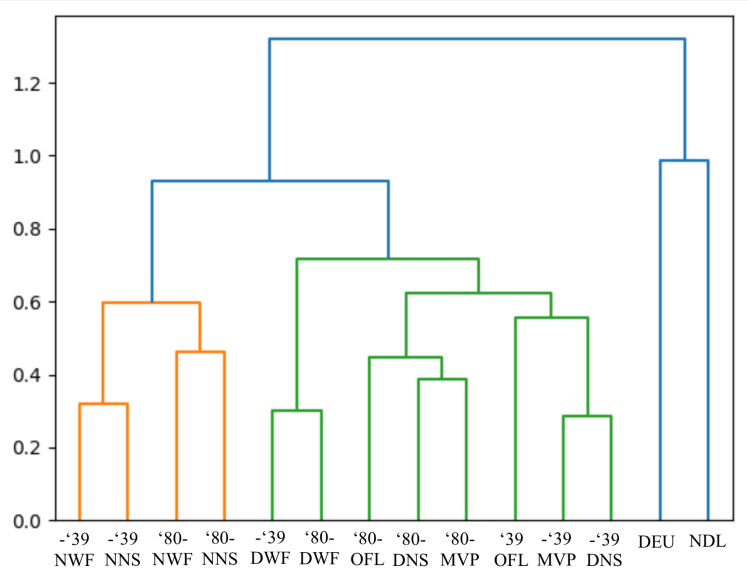
RESULTS – POS ONLY

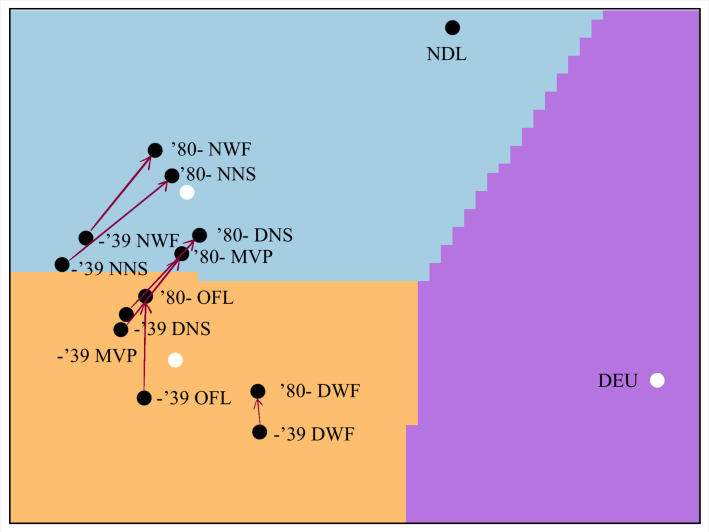




Low Saxon dialects approach the majority languages and each other.

RESULTS – POS WITH MORPH. FEATURES





Both German Low Saxon and Dutch Low Saxon converging towards Dutch?

N-GRAM DIFFERENCES

- ADP, DET bigram less frequent in Low Saxon, but frequency increases
- Dative case: high values for German and (German) Westphalian
- Gender-ambiguous pronouns and articles: high values for Dutch and Low Saxon
- Issues: E.g. ’Case=Acc,Dat’ misannotation in German

SUMMARY

- Different approaches lead to different classifications:
  - Character-based models place Dutch Low Saxon closer to Dutch than what PoS and morphological features suggest.
  - PoS and morph. features more clearly show trends of linguistic change.
- The political border does play a role for classification, but does not seem to be *the* major factor explaining the development.
- No support for the traditional east-west division.

DISCUSSION

- PoS and morph. features do not show a clear trend of Low Saxon growing apart
- Limitations:
  - Relatively low feature accuracy
  - Unbalanced amount of data
  - Different sizes of dialect regions
  - Age of writers
- Future research:
  - Lemmatisation and lexical similarity
  - Older Dutch and German data

ACCESS & LICENCE

- CC BY-NC
- Helsinki NLP: `https://github.com/Helsinki-NLP/LSDC-morph/tagging`

Thanks to Kevin Behrens, Behrend Böckmann, Johanna Bojarra, Marita Bojarra, Christiane Ehlers, Marianne Ehlers, Heiko Gauert, Jan Graf, Bernd Lubs, Christian Peplow, Karl Peplow, Gennadi Ratson, Heinrich Siefert and Florian Wille for providing additional texts in the northern dialects from Germany!

Dank jüm!
Thank you!