Chapter 8

Corpus-based Low Saxon dialectometry

Janine Siewert[a], Yves Scherrer[a,b] & Martijn Wieling[c]

[a]University of Helsinki [b]University of Oslo [c]University of Groningen

In this corpus-based study, we explore how the similarity of Low Saxon dialects among each other and to the state languages Dutch and German has changed from the 19th century to today. In particular, we want to investigate if the traditional classification into an eastern and a western group is visible in the data and if the Low Saxon dialects can be found to diverge at the Dutch-German border.

We apply principal component analysis and hierarchical clustering to n-grams of characters, part of speech tags and morphological features and observe divergent developments at the separate levels. As a reflection of different orthographic traditions, a noticeable distance between Dutch Low Saxon and German Low Saxon can be attested at the character level. At the PoS and morphological level, we however find a particular closeness between Dutch Low Saxon and the northern dialects from Germany, while we see German Westphalian in an outlier position. A shift towards the state languages can be observed at the PoS level, but the overall distance between Dutch Low Saxon and German Low Saxon does not seem to markedly increase at the three levels we studied.

1 Introduction

In the context of a research project on dialectal variation in Low Saxon[1], we investigate how the similarity of larger dialect areas has changed from the 19th to the 21st century. Our study is based on corpus data that we have collected earlier and covers both the Dutch and the German side of the Low Saxon language area.

[1] Also referred to as "Low German".

Janine Siewert, Yves Scherrer & Martijn Wieling. 2025. Corpus-based Low Saxon dialectometry. In Susanne Wagner & Ulrike Stange-Hundsdörfer (eds.), *(Dia)lects in the 21st century: Selected papers from Methods in Dialectology XVII*, 173–200. Berlin: Language Science Press. DOI: 10.5281/zenodo.15006607

Traditional dialect classifications, such as the ones presented by Schröder (2004), have large-ly relied on the occurrence of certain phonological and morphological traits. In this study, we however take a corpus-based approach and zoom in on three levels of the language: orthographic, morphological, and syntactic. Among possible differences at these three levels, we particularly want to investigate how the corpus-based results relate to the more traditional classifications and to the language contact situation. Therefore, we are interested in 1) the traditional east-west division, and 2) what the effect of the Dutch-German border is on the development of Low Saxon in recent decades.

2 Low Saxon dialect classification and variation

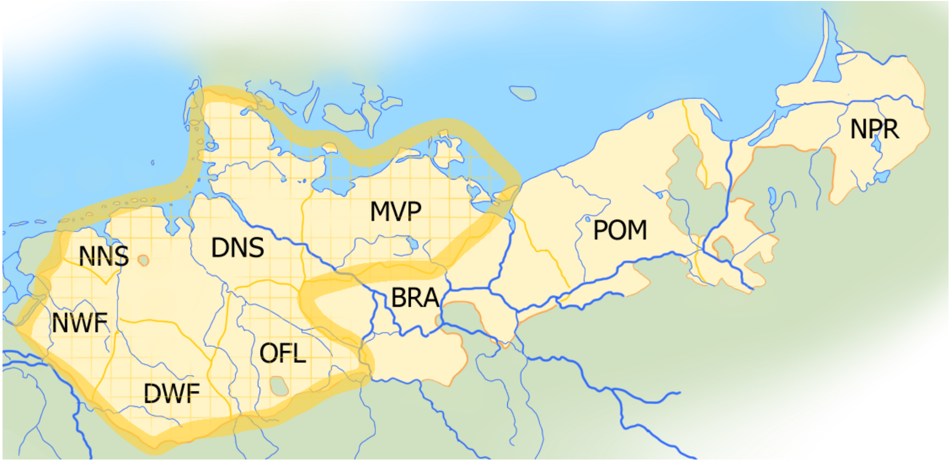

Figure 1: Low Saxon dialects: NNS: Dutch North Saxon, NWF: Dutch Westphalian, DNS: German North Saxon, DWF: German Westphalian, MVP: Mecklenburgish–West Pomeranian, OFL: Eastphalian, BRA: Brandenburgish, POM: East Pomeranian, NPR: Low Prussian; self-created map

Low Saxon is a West-Germanic language primarily spoken in northern Germany and the north-eastern Netherlands. Despite official recognition in both countries, there is no common standard variety. As a result, Low Saxon speakers will typically speak and write their own dialect(s) in all language use contexts.

Figure 1 shows the Low Saxon language area in the early 20th century with its major dialect groups. In this paper, we will focus on the encircled dialects.

The dialect division of German Low Saxon follows the traditional classification presented by Schröder (2004) and Stellmacher (1983). This classification is primarily based on phonological and morphological traits such as vowel phoneme mergers and the plural ending of verbs in the present tense. Based on this plural suffix, Stellmacher (1983: 240) divides German Low Saxon into an eastern and a western group, where *-et* is used in the west and *-en* in the east. Of the dialects included in our research, only Mecklenburgish-West Pomeranian (MVP) belongs to this eastern group.

A more fine-grained division is normally used for Dutch Low Saxon (see for instance Bloemhoff et al. 2019: 20), possibly due to the much smaller size of the language area. In this cross-border comparison, we wanted to keep the size of the areas of comparison roughly in line with one another, but at the same time be able to look at internal differences within Dutch Low Saxon. As a compromise, we opted for a two-fold North-South division. Dutch North Saxon (NNS) is only represented by texts from Groningen in our dataset, while the southern part, called "Dutch Westphalian" (NWF) here, includes texts from Achterhoeks, Veluws, Twents, Sallands and Drents. This North-South division has a rough correspondence with German North Saxon (DNS) and German Westphalian (DWF).

An important isogloss for dividing North Saxon from Westphalian is the merger of Proto-Germanic *â and lengthened *a in the north contrasting with the preservation of distinct phonemes in the Westphalian dialects (Niebaum 2008, Bloemhoff et al. 2019). This distinction can still be found on both sides of the border. In addition, the northern group of Dutch Low Saxon has a Frisian substrate which is likely why Jellinghaus (1892) refers to it as "Friso-Saxon" and contrasts it with "Saxon" dialects.

2.1 Orthographic differences

In addition to dialectal variation, we find a large variety of spelling traditions and personal writing habits. While regional language institutes or societies, publishers or individual language enthusiasts devise and spread their own spelling systems today, explicit normification was not this widespread in the 19th century. Even today, every Low Saxon speaker does not adhere to the spelling that a local publisher from the same region might prefer. For instance, while textbooks from Mecklenburg–West Pomerania (MVP) (e.g., Herrmann-Winter 2006) tend to use the spelling by Herrmann-Winter, and textbooks from several areas of German North Saxon (DNS) (e.g., Arbatzat 2016, Hiestermann & Konen-Witzel 2021) can often be seen to follow the SASS[2] spelling (Kahl & Thies 2009), some of the people who provided texts for our dataset had other spelling preferences.

[2] Not an abbreviation but based on the family name of Johannes Saß.

Local spelling traditions tend to draw inspiration from the majority language orthography to different degrees and in different ways. We will illustrate this with a few examples from our dataset. Sentence (1) is from the Netherlands and sentences (2) and (3) from Germany. This can easily be recognised by the capitalisation of nouns (e.g., *Kaarten*, *Ogen*, *Dage*, *Magister*) or lack thereof (*leu*, *weerde*, *ding*). While not all Low Saxon writers from Germany capitalise nouns, it is a common trait.

(1) De leu veult gemènlik eerst de weerde van ’n ding, as ze ’t nig hebt.
‘In general, people only become aware of the value of things when they do not have them.’

(2) 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.
‘Arfest bought two first class tickets, and when I looked on at him with astonishment, he laughed and said that may be written there but I should just get on.’

(3) Eunige Dage später frogere de Magister, biu de veuer Johrestyien herren.
‘A few days later, the teacher asked what the four seasons were called.’

Sentence (1) furthermore shows two grapheme-phoneme correspondences typical for Low Saxon writing from the Netherlands: <eu> for /ø/ and <z> for /z/. In texts from Germany, one would rather expect <ö(ö)> and <s>.

In Sentence (2) we find the letter <ß> for /s/,[3] again a strong indicator of the text coming from Germany as this letter is likely not used in Dutch Low Saxon. In addition, we see a difference in the adaptation of German spelling rules between Sentence (2) and (3). In the word *Johrestyien* ‘seasons’ (compare German *Jahreszeiten*) in (3), the <h> is used as part of a digraph marking a long vowel. According to some local spellings such as the SASS (Kahl & Thies 2009), the <h> should indeed be used as a length marker if it is written in the German cognate. However, such a principle is not followed by all Low Saxon writers in Germany as can be seen in Sentence (2). If this writer had adhered to the same rule, we would find <nehm> ‘took’ and <stahn> ‘stand’ instead of <neem> and <staan>, compare *nahm* and *stehen* in German.

[3] Here used according to the old German orthography, where the <ß> replaced <ss> in coda position.

2.2 Language contact

Another layer of variation is added by the influence from the different state languages, Dutch and German. Goossens (2019) indeed observed a replacement of certain local characteristics with features from the majority languages. He finds this to primarily affect the lexical level, but influence can by found at the phonological, morphological, and syntactic level as well.

Influence on grammatical gender, a morphological feature relevant for our current research, is described by Bloemhoff et al. (2019: 107–109). Whereas in Dutch, masculine and feminine gender have merged to a common gender, there are Dutch Low Saxon dialects such as Twents, where this distinction has been preserved. Bloemhoff et al. (2019) however find that speakers of Twents are increasingly unsure which of the two genders to assign to nouns with common gender in Dutch.

In his comparison of two neighbouring dialects at the border, Smits (2011) observes two different tendencies. While he finds more structural language loss, i.e., convergence towards the state language, on the Dutch side of the border, there is more overall language loss, i.e., language shift from Low Saxon to German, on the German side. Nevertheless, the remaining speakers on the German side retain more structural characteristics of their dialect than their neighbours from the Netherlands.

2.3 Morphological and syntactic traits

Next, we will discuss six morphological and syntactic traits where Low Saxon dialects differ from each other and/or from the respective majority language. We focus on the kind of structures that are visible at the PoS and morphological level and where we expect the majority languages to exert influence. The traits are presented in the same order in which we discuss them in Section 5.2, Section 5.3 and Section 6.

2.3.1 Lack of the definite article

According to Pheiff (2022), there are Low Saxon dialects where the usage of the definite article is less frequent than in Standard Dutch and German. He observes this to be particularly true for the north of Groningen but has attested the lack of a definite article in contexts where it would be obligatory in the majority languages in other regions as well.

For instance, our dataset contains the following sentence in older German North Saxon (DNS):

(4) As de Bur 'n Sett up Bedd wäsen weer.
when the farmer a moment on bed be.PAST-PTCP be.PST.3SG
'When the farmer had been on the bed for a moment.'

This sentence illustrates the tendency mentioned by Pheiff (2022: 147) that the definite article is missing particularly after prepositions: While no article is used after *up*, it does occur in *de Bur* and *'n Sett*. Proper nouns are another context where the usage of the definite article is less frequent. This is in line with the geographical distribution of the definite article with proper nouns within German as well, where the use of the definite article with personal names decreases the further north you go (Pheiff 2022: 150).

2.3.2 Motion verb + (to/and) + infinitive

German and Dutch combine motion verbs with the construction *um zu*/*om te* to express finality. In Low Saxon, while usage of *üm to* is possible, plain *to* similar to English occurs commonly as well:

(5) He keem na de Köök to fröhstücken.
he come.PST.3SG to the kitchen to eat_breakfast
'He came to the kitchen to eat breakfast.' (Thies 2010: 74)

In addition, in a few dialects, the infinitive can be connected with *un/en* 'and'. Thies (2010: 76) mentions this construction for Schleswig and assumes Danish influence but, for instance, Wilhelm Wisser from Eastern Holstein uses it in his collection of fairy tales as well:

(6) Hê schall mal na ehr'n Brôder Mạn gahn un fragen
he shall.3SG once to her-M.ACC brother moon go and ask
den' mal.
that.M-ACC once
'He should just go to her brother moon and ask him.' (Wisser 1921: 53)

Furthermore, this construction is attested in Dutch Low Saxon dialects such as Stellingwerfs: *Ie kun mar beter naor de zoolder* ***gaon en vang(en)*** *die moes* 'You better go to the attic and catch the mouse.' (Bloemhoff 2008: 191) and Gronings: *ik* ***goa hin en helpn*** *Kloas* 'I will go and help Kloas.' (van Bree 2008: 114).

2.3.3 Case inflection

Low Saxon dialects differ noticeably in how many cases nouns inflect for. In Medieval Low Saxon, we still find nominative, genitive, dative, and accusative (Lasch 1974), but the modern varieties typically display a simplified case system. According to Lücht (2016: 62), East Frisian[4] does not inflect nouns for case. Usage of independent genitive forms is generally very restricted, but most German Low Saxon dialects still distinguish between nominative and accusative in singular masculine nouns, which can be seen in the form of the adjective and the article (Lindow et al. 1998: 144, 191).

Many German Low Saxon dialects have preserved remnants of the dative case that might be used in connection with certain prepositions. In the German North Saxon (DNS) Sentence (7) from our dataset, after *bi*, we see a form of the definite article resembling the masculine accusative. As the expected form in the accusative of this neuter noun would be *'t*, this can be interpreted as a reduced version of the older dative form *'m*.

(7) 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.'

To is another preposition where we often encounter the definite article of neuter nouns as *'n* instead of otherwise expected *'t*, e.g., in this German North Saxon (DNS) sentence from our dataset: *Dat's to'n Lachen!* 'That is ridiculous!'

Independent productive dative forms have only been preserved in a few southern German Low Saxon dialects such as South Westphalian (part of DWF). While most northern dialects in Germany distinguish between *de Disch* 'the table' in the nominative and *den Disch* in the non-nominative, parts of German Westphalia exhibit a three-fold distinction between *de Disk*[5] in the nominative, *dem Diske* in the dative, and *den Disk* in the accusative (Lindow et al. 1998: 144–145).

Case distinctions are not commonly mentioned in descriptions of Dutch Low Saxon and neither have we encountered any in our data. Individual fossilised cases in fixed expressions similar to Dutch might however still exist.

[4] Part of German North Saxon (DNS). Not to be confused with the East Frisian language Saterland Frisian.

[5] The nominative form is not given by Lindow et al. (1998: 145), but based on the other forms they present, it likely looks like this.

2.3.4 Subjunctive

Within German Low Saxon, there is a North-South divide in subjunctive usage. As can be seen in the i-mutated forms *söl* and *bekäme* in Sentence (8) from German Westphalia, a few southern Low Saxon dialects have preserved distinct subjunctive forms.

(8) 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.' (Saltveit 1983: 299)

Productivity however decreases further north. In Sentence (9) from Schleswig-Holstein, we find no i-mutation in *schusst* although *irrealis* meaning can be deduced from the context.

(9) Du schusst man lewer to Huus gahn hebben.
you.SG shall.PST-2SG but rather to house go have
'You had better gone home.' (Saltveit 1983: 300)

The past tense forms in many northern dialects in Germany may show i-mutation due to their origin in subjunctive forms, but synchronically, they have taken on the role of the past indicative and thus there is no formal difference between indicative and subjunctive. These indistinct forms in the northern dialects can function as both indicative and subjunctive (Saltveit 1983: 298–301).

In addition, *irrealis* meaning can be expressed by means of auxiliary verbs such as *willen* 'to want', *warden* 'to become', *schölen* 'shall' and *doon* 'to do', (cf. Lindow et al. 1998, who mainly describe German North Saxon (DNS)).

We are not aware of the existence of distinct subjunctive forms in today's Dutch Low Saxon.

2.3.5 Infinitivus pro participio (IPP)

Infinitivus pro participio refers to the phenomenon in West-Germanic languages such as Dutch and German of using an infinitive instead of an expected past participle, for example:

(10) Ich hätte das tun können.
I have.PST.SBJV-1SG that do.INF can.INF
'I could have done that.'

Schmid (2005: 1) lists "Low German"[6] as one of the West Germanic languages where IPP-constructions do not appear, which is in line with Lindow et al. (1998).

(11) Korl hett den Text nich lesen kunnt.
Korl has the-ACC.SG text not read can.PST-PTCP
'Korl could not read the text.' (Lindow et al. 1998: 108)

Bloemhoff et al. (2019: 66) present a more varied picture for Dutch Low Saxon. They state that the northern dialects Gronings and Stellingwerfs indeed do not know the IPP-construction, whereas, in the other Dutch Low Saxon dialects, they assume a correlation with presence or absence of the *(g)e*-prefix in the past participle.

The situation for German Low Saxon is also in fact more complex than presented above. Even in the north-western dialects on which the grammar by Lindow et al. (1998) focuses, some speakers today do use the IPP-construction (personal observation), perhaps due to influence from Standard German.

2.3.6 Complementiser doubling in subordinate clauses

Similar to Frisian (Popkema 2018: 299), but different from (Standard) German and Dutch, *as* 'as' or *dat* 'that' can occur as a second complementiser in subordinate clauses, typically after question words. This is well attested in Dutch Low Saxon, cf. the following example from Gronings:

(12) Ik mout waitn wel dat ik in hoes krieg.
I must.1SG know who that I in house get.1SG
'I need to know whom I will get into the house.' (van Bree 2008: 114)

This phenomenon is not commonly described in grammars for German Low Saxon but Saltveit (1983: 289, 330) briefly mentions the usage of both *dat* and *as* and offers two example sentences. Moreover, we have encountered several examples with *as* in literary works such as Wisser (1921) and Peters (1986).

(13) Un dạrmit secht de ol Mann em Beschêd, wodenni as he dat
and therewith says the old man him information how as he that
maken schall.
make shall.3SG
'And with this, the old man tells him how he should do it.' (Wisser 1921: 29)

[6] Due to the name choice possibly only referring to varieties from Germany.

Schallert et al. (2018) have found attestations of the variant with *dat* already in Medieval Low Saxon. Its usage is however not restricted to Low Saxon, but they point out that this type of construction occurs throughout West-Germanic and even neighbouring Romance and Slavonic varieties, in particular in the Alpine region.

3 Dataset

Our dataset is taken from the PoS-tagged and morphologically annotated version[7] of the LSDC dataset LSDC-morph (Siewert et al. 2022).

The overall dataset covers eight dialect regions from the 19th, 20th and 21st century, but in this study, we focus on these six major Low Saxon dialect groups: Dutch North Saxon (NNS), German North Saxon (DNS), Dutch Westphalian (NWF), German Westphalian (DWF), Eastphalian (OFL) and Mecklenburgish-West Pom-eranian (MVP). The reason for excluding the other eastern dialects Brandenburgish (BRA), East Pomerian (POM) and Low Prussian (NPR) is the relatively low amount of data.

The older part of the dataset consists primarily of copyright-free material available online, mainly on Wikisource, Leopold & Leopold (1882)[8] and the *Twentse Taalbank* (van der Vliet 2021). On the other hand, the modern data for most dialects was personally provided by a variety of local authors. Common genres in the dataset are short stories and short novels, but various other genres such as speeches, religious texts, historical accounts, fairy tales and letters are included as well. For a more detailed description of the content and data collection for the LSDC dataset, please see Siewert et al. (2020).

We split the data into two time periods: 1800–1939 and 1980–today[9]. This split is motivated by the language shift to the respective majority language, which in most regions occurred between the 1940s and 1980s. A practical reason possibly connected to the language shift is the lack of data from the intermediate period. Only from German Westphalian (DWF) we have one text that was marked as published during this period and three additional short texts that might have been published then based on the authors' life dates. Another practical expla-

[7] Please see https://universaldependencies.org/u/pos/index.html and https://universaldependencies.org/u/feat/index.html for a description of the PoS tags and morphological features.

[8] Digitised by dbnl: https://dbnl.nl/tekst/leop008sche00_01/.

[9] In practice, this likely means roughly 2000–today, but we do not have the exact year of publication/writing of every text that local authors provided.

nation for the lack of data might be that texts from this period are often still copyright-protected and therefore not easily available in digitised format.

We thus have an older period, when Low Saxon was still the dominant language of oral communication, and a newer period, when (a regional version of) Dutch or German has become the primary language in everyday life.

Our Standard German (DEU) and Standard Dutch (NDL) datasets are taken from Universal Dependencies.[10] For German, we used the UD_German_HDT treebank (Borges Völker et al. 2019) and, for Dutch, both the UD_Dutch-Alpino and the UD_Dutch-LassySmall treebanks (Bouma & van Noord 2017). We filtered the Dutch and German treebanks to remove duplicate sentences.

The overall size of our dataset is shown in Table 1 and our updated version of LSDC-morph dataset is made available at https://github.com/Helsinki-NLP/LSDC-morph/tree/main/methods2022.

Table 1: Size of the dataset

	1800–1939		1980–2022	
	Sentences	Tokens	Sentences	Tokens
Dutch North Saxon (NNS)	1,828	44,067	17,796	261,342
Dutch Westphalian (NWF)	4,948	102,656	9,245	136,992
German North Saxon (DNS)	23,075	429,122	3,526	61,174
Mecklenburgish–West Pomeranian (MVP)	20,007	577,767	3,055	42,434
German Westphalian (DWF)	17,450	337,871	16,015	273,513
Eastphalian (OFL)	1,684	33,162	8,441	145,792
Standard German (DEU)			185,380	3,489,305
Standard Dutch (NDL)			20,591	306,028

3.1 Annotation: Changes made to the Dutch and German UD datasets

Due to the differences in morphological feature annotation of the UD datasets, reannotation of the Standard Dutch (NDL) and Standard German (DEU) data was necessary. For this, we manually modified the annotation of 200 Dutch and German sentences each in order to train annotation models on these. For Dutch, we mainly extended morphological feature annotation, added e.g., PronType, full

[10] https://universaldependencies.org

marking of person and number for verbs. There were only marginal adaptations for PoS, specifically relating to what is considered a proper noun or a number. For German, we added the differentiation for PronType, we added case marking to nouns (in the original sometimes only marked on the article), and we removed the case label from proper nouns unless they actually were marked.

3.2 Tagging

We have made substantial manual corrections to the automatic tokenisation of the LSDC-morph dataset. In addition to wrong or missing sentence splitting after certain punctuation marks, this mainly concerned different orthographic solutions for contractions. For instance, 'on the-M.SG' might be realised in the form of *up'n*, *up n* or *upm* among others, which leads to three different realisations at the PoS level: 'ADP – PUNCT – DET', 'ADP – DET' and 'ADP'. These were unified to the 'ADP – DET' format.

Due to the corrections to the tokenisation, the Low Saxon data needed to be re-annotated just like the Dutch and German data. The automatic tagging was done with the Stanza tagger (Qi et al. 2020)[11] pretrained on large datasets without target annotation and fine-tuned separately for German, Dutch, Dutch Low Saxon (NNS and NWF), northern German Low Saxon (DNS and MVP) and southern German Low Saxon (DWF and OFL) on small sets of manually annotated data. This split of Low Saxon into three training groups was motivated by the increase in performance observed in unrelated lemmatisation and PoS tagging experiments done by us earlier (unpublished).

For training the Stanza tagger, we used the pretrained embeddings from the CoNLL 2017 shared task[12] for Standard Dutch and Standard German and, for Low Saxon, the fastText embeddings by Grave et al. (2018). In addition, we trained our own Low Saxon embeddings on our dataset using GloVe (Pennington et al. 2014). These led to a small improvement in accuracy for Dutch Low Saxon (91% to 92% for PoS and 80% to 83% for features) compared with previous training using fastText embeddings. So, presumably, the much larger fastText embeddings mostly or only represent German Low Saxon. As we furthermore receive a noticeably better model accuracy for northern dialects (96% for PoS and 85% for features) compared with southern German Low Saxon (91% for PoS, 83% for features), despite more southern training data, the fastText embeddings may have been mostly trained on northern German Low Saxon data.

[11] The stand-alone version we used is available at https://github.com/yvesscherrer/stanzatagger.

[12] Available at https://stanfordnlp.github.io/stanza/word_vectors.html

4 Approaches

4.1 Data encoding

The three levels under investigation are represented by bigrams and trigrams. We removed n-grams that occurred five times or less as well as n-grams that contained the tags ‘SYM’, ‘_’, ‘X’ or two consecutive ‘punct’ tags. The motivation for this is that the tags ‘SYM’, ‘_’ and ‘X’ do not represent linguistic elements of interest, and that we did not want to give too much weight to personal writing habits, such as the use of doubled quotation marks by certain 19th century authors.

Character n-grams (68,695,124 overall n-grams, 30,762 distinct ones) approximate the orthographic level but can be assumed to also capture phonological and morphological features. PoS (Part of Speech) n-grams (10,811,793 overall n-grams, 2,794 distinct ones) match the syntactic level, and n-grams of PoS and morphological features (10,207,431 overall n-grams, 68,551 distinct ones) correspond to morphology and morpho-syntax. Table 2 shows character and PoS n-grams that we extract from the older Dutch Westphalian sentence from our dataset *Met un moand!* ‘With(in) a month!’.

Table 2: Extracted n-grams

	bigrams	trigrams
characters	(‘m’, ‘e’), (‘e’, ‘t’), (‘t’, ‘ ’), (‘ ’, ‘u’), (‘u’, ‘n’), ...	(‘m’, ‘e’, ‘t’), (‘e’, ‘t’, ‘ ’), (‘t’, ‘ ’, ‘u’), (‘ ’, ‘u’, ‘n’), ...
PoS	(adp, det), (det, noun), (noun, punct)	(adp, det, noun), (det, noun, punct)

The first trigram of the same sentence combining PoS and morphological features is:

```
(ADP, AdpType=Prep),
(PRON, Case=Acc,Dat|Definite=Ind|Gender=Masc|Number=Sing|PronType=
    Art),
(NOUN, Case=Acc,Dat|Gender=Masc|Number=Sing).
```

In these n-grams, the first part of each element is the PoS tag and the second part consists of a concatenation of the morphological features. In dialects that have lost certain distinctions, such as the accusative and dative distinction in case of the one above, we use combined values, here `Case=Acc,Dat`[13].

N-grams have been shown to constitute a suitable unit for corpus-based dialectometry. Wolk & Szmrecsanyi (2016) compared the performance on PoS n-grams to manually selected features in their study of British dialects and discover that PoS n-grams lead to a comparable performance. In their Swiss German dialect identification experiments, Malmasi & Zampieri (2017) discover that character n-grams outperform word-based ones.

4.2 Dialectometry background

Dialectometry is a branch of dialectology where quantitative, and today commonly computational, methods are used in order to measure the difference between language varieties. In contrast with more traditional dialectological approaches, dialectometry typically makes use of a large aggregate of features instead of individual or a small number of selected features (Wieling & Nerbonne 2015). Nerbonne (2009) stresses the advantage of aggregate similarity as manually selected features are prone to cherry-picking.

In his PhD research, Spruit (2008) used quantitative approaches to investigate syntactic variation in West Germanic varieties spoken in the Netherlands and Flanders. He also compared the syntactic level with pronunciational and lexical differences, and was able to show that while a certain correlation can be found between the different levels, there is no full overlap. E.g., whereas Frisian clearly stands out at the lexical level, it appears fairly similar to the Low Saxon varieties at the syntactic level. (Spruit 2008: 75–78)

From the perspective of our current research, the reanalysis of the Wenker language atlas data by Lameli (2016) is particularly interesting. Based on linguistic variables, he calculates the aggregate similarity between different locations and uses hierarchical clustering to identify larger German Low Saxon dialect regions. The result are three major dialect areas: Northern Low Saxon (North Saxon and Mecklenburgish–West Pomeranian), Southern Low Saxon (Westphalian and Eastphalian) and Brandenburgish. While most of the areas found correspond roughly to the traditional dialect regions, the north-western part[14] of what is traditionally considered Westphalian here clusters with the northern group. This is especially important as it concerns the border region to the Netherlands. If this north-western part of German Westphalian at the Dutch border in fact shares

[13] The nominative value is, however, always kept separate, because all dialects have preserved this distinction at least in the personal pronouns.

[14] The German Westphalian data in our dataset mostly come from the southern and eastern part.

more features with the northern dialects, one might also expect the Dutch Westphalian dialects to resemble the northern group more closely.

4.3 PCA and hierarchical clustering

In the experiments presented below, we make use of PCA (Principal Component Analysis) with two dimensions as well as hierarchical clustering with Ward linkage and Euclidean metric. We compute the dialect distances with the help of the scikit-learn Python library (Pedregosa et al. 2011). The input to the PCA and the hierarchical clustering is a matrix of n-gram counts per variety, such as older (1800–1939) Dutch Westphalian (NWF) or contemporary (1980–2022) Eastphalian (OFL). Before clustering, the n-gram counts are tf-idf-normalised, since raw counts are not comparable due to the differences in amount of data per variety. In our context, the abbreviation can be understood as "term frequency – inverse *dialect* frequency".

We tried out different splits of the dataset to test the dialect-internal consistency and found that the different parts of the same variety indeed form clusters. Neither did the inclusion of Mecklenburgish–West Pomeranian (MVP), the addition of new data in German North Saxon (DNS) nor the reduction of the German (DEU) data to 20,000 sentences cause a noticeable effect on the overall tendencies. Furthermore, we reran the clusterings several times to test their stability and found no major differences between the individual runs. Only in the k-means clustering[15] performed on the PCA-reduced data, we found slight variation in the position of cluster borders next to or between very close varieties.

5 Results

We will first present the results of character-based experiments, followed by the PoS-based ones and, finally, the results based on both PoS-tags and morphological features. The figures show the PCA-based results on the left and hierarchical clustering results on the right. The red arrows in the PCA figure indicate the development of a particular Low Saxon dialect from the older period, 19th to early 20th century, to the more modern period, late 20th to 21st century.

In addition to the figures, we also discuss n-gram counts relating to particular phenomena. Due to their size, we cannot present the tables here but, the files can be found on GitHub: https://github.com/Helsinki-NLP/LSDC-morph/tree/main/methods2022.

[15] Not further discussed in this paper.

5.1 Character level

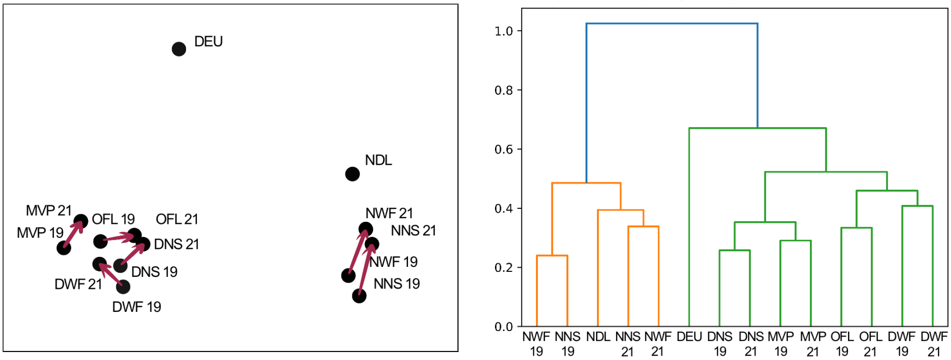

Figure 2: Character level PCA and hierarchical clustering. (NNS: Dutch North Saxon, NWF: Dutch Westphalian, DNS: German North Saxon, DWF: German Westphalian, MVP: Mecklenburgish–West Pomeranian, OFL: Eastphalian)

The character-based PCA results on the left in Figure 2 show a division into three groups: German Low Saxon (DNS, DWF, MVP, OFL), German (DEU), and Dutch (NDL) with Dutch Low Saxon (NNS, NWF). While the Dutch Low Saxon dialects seem to approach Dutch, German Low Saxon does not show a clear trend of converging towards German.

On the right, we see a division according to state, likely reflecting different orthographic traditions. Interestingly, the variants from the Netherlands cluster according to century, with older Dutch Low Saxon in one group and late 20th to 21st century Dutch and Dutch Low Saxon in the other. The variants from the German side however, branch according to what we would expect based on the analysis by Lameli (2016), with a northern (DNS, MVP) and a southern (DWF, OFL) Low Saxon branch that subsequently divide into the major dialect groups. In Dutch Low Saxon, we thus find diachronic differences more strongly pronounced, which can also be seen in the PCA. In German Low Saxon, in contrast, dialectal differences appear more important than diachronic ones.

5.2 PoS level

In Figure 3, visualising the PoS results, we see Dutch (NDL) and German (DEU) forming a common group and the Low Saxon dialects forming another one. Particularly the PCA results show the Low Saxon dialects approaching the modern

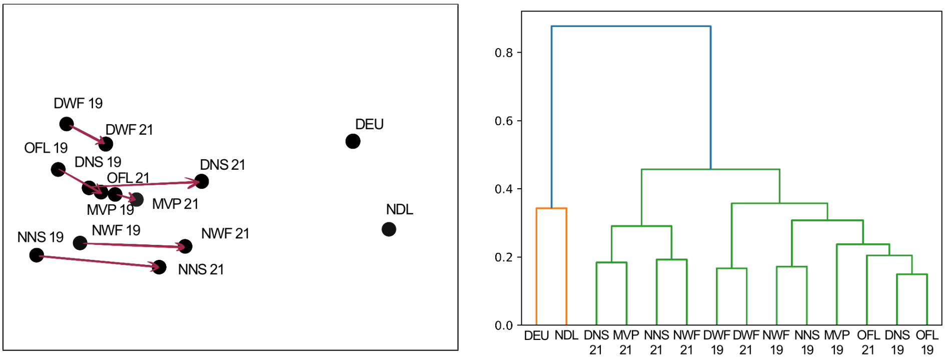

Figure 3: PoS level PCA and hierarchical clustering.
NNS: Dutch North Saxon, NWF: Dutch Westphalian, DNS: German North Saxon, DWF: German Westphalian, MVP: Mecklenburgish–West Pomeranian, OFL: Eastphalian

state languages Dutch and German. This does, however, not increase the distance along the border in all cases. For example, contemporary German North Saxon (DNS 21) is not placed closer to German Westphalian (DWF) than to the contemporary dialects from the Dutch side (NNS 21, NWF 21).

In the hierarchical clustering, we find a somewhat corresponding constellation where the contemporary northern branch from Germany (DNS 21, MVP 21) forms a cluster with contemporary Dutch Low Saxon (NNS 21, NWF 21), while the remaining Low Saxon dialects form another cluster. Within this second cluster, we find German Westphalian (DWF) branching off first while older Dutch Low Saxon (NNS 19, NWF 19) is placed closer to Eastphalian (OFL) and the older variants of the northern branch of German Low Saxon (DNS 19, MVP 19).

In the use or lack of the article described in Section 2.3.1, we indeed see similar trends to what Pheiff (2022) described. For the n-gram (‘ADP’, ‘DET’, ‘NOUN’), we find the lowest score in older Dutch North Saxon (NNS 19, NWF 19). However, comparing the older part of the dataset to the contemporary one, we can observe the score to increase for all dialects except Mecklenburgish–West Pomeranian (MVP). Similarly, for the n-gram (‘ADP’, ‘DET’, ‘ADJ’), all Low Saxon variants – both older and contemporary – receive a score lower than Dutch (NDL) and German (DEU), and, again apart from Mecklenburgish–West Pomeranian (MVP), show an increase towards the modern period. Also, the scores for the n-gram (‘DET’, ‘PROPN’) show lower values for all Low Saxon varieties compared with Dutch (NDL) and German (DEU). Again, an increase can be attested in all varieties except one. A slight decrease can be seen in German North Saxon (DNS).

Dialect differences can also be found for the infinitive construction with *üm*, cf. Section 2.3.2, represented by the n-gram ('SCONJ', 'PART', 'VERB').[16] We do not find this n-gram in the German North Saxon (DNS) or Mecklenburgish–West Pomeranian (MVP) part of the dataset. Neither do we find it in the older Dutch North Saxon (NNS 19, NWF 19) or older Eastphalian (OFL 19), but this could be due to the small amount of data in these two varieties (cf. Table 1) and the relative rarity of the construction, as it does appear in the larger contemporary dataset. German Westphalian (DWF) receives comparable scores to German (DEU) (0.00015) in both parts of the dataset (0.00011 and 0.00014). Dutch Westphalian (NWF), on the other hand, shows a decrease from 0.00093 to 0.00015, a score similar to German (DEU) and German Westphalian (DWF).

At the PoS level, we cannot observe a clear tendency of Low Saxon growing apart at the political border. While the increase in article usage brings the dialects closer to Standard Dutch (NDL) and German (DEU), this same development occurs on both sides of the border. Instead, we find a particular closeness of Dutch Low Saxon (NNS, NWF) to the northern dialects (DNS, MVP) in Germany and to a somewhat lesser degree to Eastphalian (OFL).

5.3 PoS and morphological features

In Figure 4, the distance to Dutch (NDL) is disproportionally increased by the gender feature. Since Dutch does not distinguish between masculine and feminine gender in nouns anymore, these receive the tag `Gender=Fem,Masc`, while Low Saxon and German nouns mostly receive the distinct gender features `Gender=Fem` or `Gender=Masc`. Compare Figure 5, where the *masculine* and *feminine* value of this feature have been replaced by *com* in the whole dataset. This does, however, not seem to obscure the general trends of development within Low Saxon.

An unexpected finding in the PCA is that not only Dutch Low Saxon (NNS, NWF), but also the northern branch of German Low Saxon (DNS, MVP) converges towards Dutch (NDL). Eastphalian (OFL) on the other hand, appears to approach the northern dialects. In the hierarchical clustering, similar to Figure 3, Dutch Low Saxon forms a cluster with Eastphalian and northern German Low Saxon.

We find German Westphalian (DWF) in an outlier position both in the PCA and in the hierarchical clustering. In Figure 5, it can even be seen to cluster with German (DEU) instead of the other Low Saxon dialects. In addition to the

[16] Although this does not cover cases with an object or adverb between the *üm* and the *to*, it can still give us a rough idea of the usage.

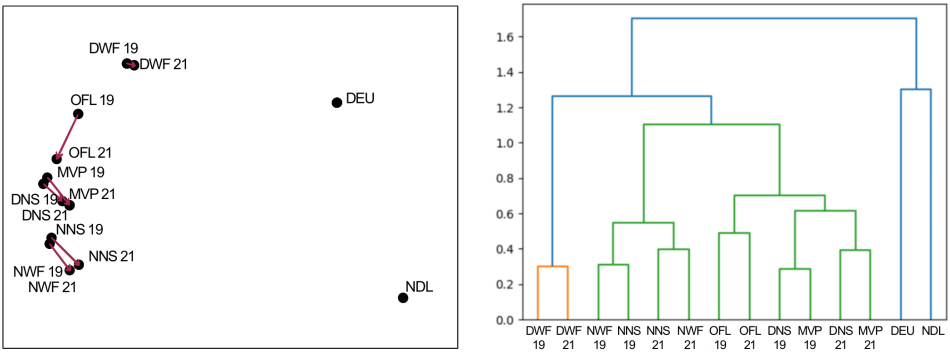

Figure 4: PoS and morphological features, PCA and hierarchical clustering.
NNS: Dutch North Saxon, NWF: Dutch Westphalian, DNS: German North Saxon, DWF: German Westphalian, MVP: Mecklenburgish–West Pomeranian, OFL: Eastphalian

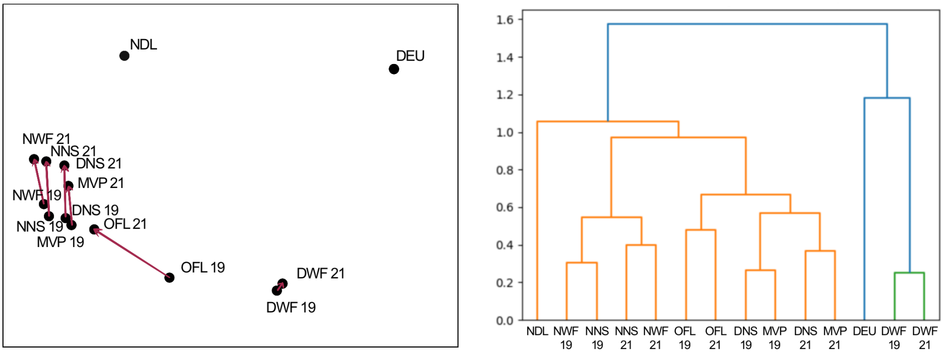

Figure 5: PoS and morphological features, PCA and hierarchical clustering, without feminine–masculine distinction
NNS: Dutch North Saxon, NWF: Dutch Westphalian, DNS: German North Saxon, DWF: German Westphalian, MVP: Mecklenburgish–West Pomeranian, OFL: Eastphalian

greater closeness to German, the amount of change in German Westphalian is also strikingly small. Preservation of the dative case, cf. Section 2.3.3, is certainly one feature that contributes to the relative closeness to German. Among the n-grams with the highest scores for both German and German Westphalian, there are several ones containing the `Case=Dat` feature.

The subjunctive mood described in Section 2.3.4 was unfortunately not learnt properly by the tagger due to its apparent rarity in the training data. As subjunctive forms were tagged as indicative or ambiguous, we cannot draw any conclusions based on our dataset regarding this aspect.

The IPP construction[17] or lack thereof[18] described in Section 2.3.5 is not among the highest scoring features, as only the bigram

```
('AUX', 'VerbForm=Inf'), ('VERB', 'VerbForm=Inf')
```

makes it into the top 500 features for older Dutch Westphalian (NWF) and Dutch (NDL), whereas in most cases, the relevant n-grams are not even among the top 1,000 features. Nonetheless, we do observe dialectal differences, as well as language change here. These, however, only partly reflect the description in the literature. The main finding is that the IPP construction, either with *VERB – AUX* or with *AUX – VERB* order, occurs in all Low Saxon dialects and is the preferred construction in most of them. Curiously, it is by far the preferred construction even in both periods of Dutch North Saxon (NNS), while Bloemhoff et al. (2019) claim to not know of the phenomenon. The only varieties showing a clear preference for the non-IPP version, are German North Saxon (DNS) and Mecklenburgish–West Pomeranian (MVP) from the older period. In the contemporary data, despite the non-IPP construction still occurring frequently, its IPP alternative has started to dominate. This confirms our personal observation that the IPP variant is indeed used as well by Low Saxon speakers today.

Despite such individual opposing trends, however, none of the German Low Saxon dialects seem to as a whole increasingly resemble German (DEU), and the distance between Dutch Low Saxon (NNS, NWF) and the northern dialects from Germany (DNS, MVP) remains roughly the same.

6 Discussion

As expected, the three levels under investigation lead to rather different results, in terms of both clustering and observable changes in similarity.

[17] Represented by the bigrams `('AUX', 'VerbForm=Inf'), ('VERB', 'VerbForm=Inf')` and `('VERB', 'VerbForm=Inf'), ('AUX', 'VerbForm=Inf')`.

[18] Represented by `('VERB', 'VerbForm=Inf'), ('AUX', 'Tense=Past|VerbForm=Part')`.

We assume the character level to largely reflect spelling differences (cf. Section 2.1) and these are likely behind the noticeable divide between Dutch Low Saxon and German Low Saxon which we do not see in the other experiments. Nevertheless, among the n-grams with the highest tf-idf values, we also find a few that indicate phonological and morphological differences. For instance, we find that the Low Saxon dialects with preserved word-final *e* (DWF, NWF, OFL) receive a higher score for the n-gram ('e', ' ') than the dialects with *e*-apocope in nouns and verbs (DNS, NNS, MVP). Similarly, the highest score for the n-gram ('t', ' ') is found in contemporary German North Saxon (DNS), which uses *-(e)t* as the plural suffix of present tense verbs, while older Mecklenburgish–West Pomeranian (MVP), which typically uses *-(e)n* for the same function, receives the lowest score.

The similarity at the character level is likely distorted by the different orthographic traditions in other ways as well. It could for example be the case that the spellings commonly used for Dutch Low Saxon today make it appear disproportionally Dutch compared with the older variants. This would explain the noticeable shift towards Dutch in the PCA as well as the grouping of contemporary Dutch Low Saxon with Dutch in the hierarchical clustering. Dutch Low Saxon clustering according to time period instead of dialect should however not primarily be attributed to orthography, as we see the same phenomenon at the level of PoS and morphological features as well.

One might have expected a comparable closeness between German Low Saxon and German due to the German-based writing traditions. In addition to morphological differences, an explanation might be found in the High-German consonant shift leading to a different distribution and frequency of many phonemes. The overall clustering within German Low Saxon corresponds well to the findings of Lameli (2016), and we see a northern group consisting of German North Saxon (DNS) and Mecklenburgish–West Pomeranian (MVP), and a southern group consisting of Westphalian (DWF) and Eastphalian (OFL).

Despite the common naming, Dutch Westphalian (NWF) could not be shown to be particularly close to German Westphalian (DWF) at any of the levels investigated. Furthermore, the similarity between Dutch North Saxon (NNS) and Dutch Westphalian (NWF) appears to be roughly equal to the similarity between German North Saxon (DNS) and Mecklenburgish–West Pomeranian (MVP). This supports the idea that if, in line with Lameli (2016), the north-western part of German Westphalian shares more features with the northern dialects than with the other Westphalian varieties, the neighbouring varieties on the Dutch side of the border might also share more traits with northern Low Saxon.

The two lines we mainly wanted to investigate – the Dutch-German border and the traditional East-West division in German Low Saxon – did not present themselves as clearly as we had expected.

An East-West division could not be found at any of the levels under scrutiny. On the contrary, German North Saxon (DNS) and Mecklenburgish–West Pomeranian (MVP) actually appear so close that they cluster according to century instead of dialect at both the PoS and the morphological level. It would be desirable to repeat these experiments with additional eastern dialects when more data becomes available.

A certain distance between Dutch Low Saxon (NNS, NWF) and the northern dialects from Germany (DNS, MVP) is observable and might be attributed to influence from the different state languages. The distance is however not bigger than between the northern German Low Saxon dialects and German Westphalian (DWF). In particular, the border does not explain why the northern dialects from Germany would converge towards Dutch (NDL) instead of German (DEU). Thus, the situation appears to be more complex.

On the one hand, the loss of morphological complexity, especially the loss of case inflection, could be one factor behind the increasing similarity between Dutch (NDL) and northern German Low Saxon (DNS, MVP). Another factor that should be taken into consideration is the ongoing strife to codify Low Saxon, in particular in Germany, where textbooks and grammar descriptions for school and adult education have been published in several dialects.

Due to this available documentation, the people who produce written German Low Saxon are probably very aware of the differences between Low Saxon and German. Moreover, they presumably strive to produce what they consider "good Low Saxon", which is not necessarily the same form of language they would informally speak at home with family and friends but might be an idealised form of their dialect where features distinct from German are preferred.

Furthermore, thanks to the internet and social media in particular, it has become easier to access content in other dialects and from the other side of the border. We are aware that several people who provided data have interdialectal and some even cross-border contact.

As a result from the larger number of speakers, northern German Low Saxon (DNS, MVP) is better represented in media and literature, which could explain why Eastphalian (OFL) appears to be converging towards the northern dialects.

Language skills might play a role as well, as the respective state language is not necessarily the only additional language of which Low Saxon writers have knowledge. Some of the people from whom we received texts are to different degrees proficient in the other state language or in Scandinavian languages as well.

Especially in case of younger speakers, knowledge of English can be assumed and English influence on the Low Saxon of younger second language or heritage learners might be a topic worth investigating.

7 Limitations and future research

We have so far based our comparisons of Low Saxon dialect similarity only on PCA and hierarchical clustering. We can therefore not preclude that other clustering approaches or a different implementation of the same clustering methods might yield different results. Furthermore, we would like to test a greater variety of visualisation techniques. For instance, a map format might be easier to grasp, especially for readers less familiar with the Low Saxon language area.

Another factor of uncertainty is the automatic tagging that our work relies on. Upon manual inspection of the n-grams, we have become aware of issues with the tagging of certain features. One of these is the gender feature. Even in the dialects that have preserved a three-fold distinction, the form of a possible accompanying article or adjective does not necessarily fully disambiguate the gender of the noun. This poses a challenge for both the automatic tagger and for the human annotator as the gender of nouns can differ from dialect to dialect and might not be documented for the precise dialect in question. In particular, the Dutch Low Saxon dialects in the process of losing their feminine-masculine distinction pose problems as it is not always obvious when a gender distinction can still be assumed.

The unfortunate fact that our tagger did not learn to recognise the subjunctive has shown that rare features can require careful manual selection of the training data. In the German and Low Saxon sentences randomly selected for manual annotation and subsequent fine-tuning, this feature apparently did not occur frequently enough.

Another tagging-related risk concerns the finetuning to different groups of dialects. Since the finetuning sets consist of only around 200 to 300 sentences per group, some overfitting to a particular subset of the data might have happened. Nevertheless, we considered the finetuning justified due to the substantial spelling variation described in Section 2.1. There are, for instance, several cases of character strings that would receive different PoS tags in different dialects: For instance, *doe* as the personal pronoun of the second person singular in Gronings (part of NNS) receives the tag 'PRON'; as the definite article in East Westphalian (part of DWF) it is tagged as 'DET' and, in addition, it is a possible spelling for the 1st person singular present tense of the verb *doon* 'to do' in, e.g., German North

Saxon (DNS) and Dutch Westphalian (NWF), when it should be tagged as ‘VERB’ or possibly ‘AUX’. Neither can we rule out that this finetuning might cause the varieties within the same group to appear closer than they are. While the members of the same group do at least not appear artificially close, compare, for instance, the German South Low Saxon group consisting of Eastphalian (OFL) and German Westphalian (DWF) in Figures 4 and 5 or the two periods of Dutch Low Saxon: Dutch North Saxon (NNS) and Dutch Westphalian (NWF) in Figure 3, larger and more diverse finetuning sets would certainly be desirable.

The different size of the dialect regions should not be neglected either. The German Low Saxon regions are noticeably larger than the Dutch Low Saxon ones, and most of the German Low Saxon texts in our dataset are not from areas particularly close to the border. Several of the German North Saxon (DNS) writers come from Schleswig-Holstein or Hamburg, while in the German Westphalian dataset (DWF), East Westphalian and South Westphalian are overrepresented. Varieties from border regions, such as East Frisian and West Munsterlandic, might have exhibited greater similarity with Dutch Low Saxon.

The lack of diversity among the writers is a problem in some dialects as well, in particular in contemporary German Westphalian (DWF) and Eastphalian (OFL), and in older Mecklenburgish–West Pomeranian (MVP). Contemporary Eastphalian is unfortunately only represented by one writer so far, which is why some of the developments we observe might simply be features of this writer’s idiolect. While in contemporary German Westphalian, we have obtained texts from a variety of writers, these mostly represent the older generation born in the first half of the 20th century. Even though the texts themselves were published in the late 20th or the 21st century, their language might be more representative of the older period. This would explain why we see so little change in German Westphalian both at the PoS and the morphological level. The older Mecklenburgish–West Pomeranian part of the dataset also includes texts from a variety of sources but works by Fritz Reuter clearly dominate. The fact that the older part of the Mecklenburgish–West Pomeranian data scores higher in article use than the newer part, could thus possibly result from characteristics of Reuter’s idiolect. The collection of additional data for these dialects would therefore be desirable.

An additional desideratum is the comparison with Dutch and German from the 19th – early 20th century. While the convergence towards the state languages that we see at the PoS level seems intuitive, we cannot yet rule out the possibility that all three languages develop into a similar direction.

Some structures such as the double complementiser with *as/dat* described in Section 2.3.6 that we initially planned to include, turned out to require lemma information since the PoS bigram *ADV – SCONJ* representing, e.g., *worüm dat* ‘why’, captures several unrelated constructions as well. We are already working on automatic lemmatisation for Low Saxon and are planning to study lexical differences as well in our future work.

Abbreviations

MMM	German	NDL	Dutch
DEU	German	NNS	Dutch North Saxon
DNS	German North Saxon	NWF	Dutch Westphalian
DWF	German Westphalian	OFL	Eastphalian
MVP	Mecklenburgish-West Pomeranian	PoS	Part of Speech

Acknowledgements

We would like to thank 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!

This work has been supported by the Academy of Finland through project No. 342859 “CorCoDial – Corpus-based computational dialectology”.

References

Arbatzat, Hartmut. 2016. *Platt: Dat Lehrbook*. Hamburg: Quickborn-Verlag.

Bloemhoff, Henk. 2008. Stellingwerfs. In Henk Bloemhoff, Jurjen van der Kooi, Hermann Niebaum & Siemon Reker (eds.), *Handboek Nedersaksische taal- en letterkunde*, 175–193. Assen: Koninklijke Van Gorcum.

Bloemhoff, Henk, Philomène Bloemhoff-de Bruijn, Jan Nijen Twilhaar, Henk Nijkeuter & Harrie Scholtmeijer. 2019. *Nedersaksisch in een notendop: Inleiding in de Nedersaksische taal en literatuur*. Assen: Koninklijke Van Gorcum.

Borges Völker, Emanuel, Maximilian Wendt, Felix Hennig & Arne Köhn. 2019. HDT-UD: A very large Universal Dependencies Treebank for German. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, 46–57. Paris, France: Association for Computational Linguistics. DOI: 10.18653/v1/W19-8006. https://aclanthology.org/W19-8006.

Bouma, Gosse & Gertjan van Noord. 2017. Increasing return on annotation investment: The automatic construction of a Universal Dependency treebank for Dutch. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, 19–26. Gothenburg, Sweden: Association for Computational Linguistics. https://aclanthology.org/W17-0403.

Goossens, Jan. 2019. „Dialektverfall“ und „Mundartrenaissance“ in Westniederdeutschland und im Osten der Niederlande. In Gerhard Stickel (ed.), *Varietäten des Deutschen: Regional- und Umgangssprachen*, 399–404. Berlin: de Gruyter. DOI: 10.1515/9783110622560-023.

Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin & Tomas Mikolov. 2018. Learning word vectors for 157 languages. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis & Takenobu Tokunaga (eds.), *Proceedings of the eleventh international conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). https://aclanthology.org/L18-1550/.

Herrmann-Winter, Renate. 2006. *Hör- und Lernbuch für das Plattdeutsche*. Rostock: Hinstorff.

Hiestermann, Heike & Katrin Konen-Witzel. 2021. *Snacken, Proten, Kören: Plattdüütsch-Lehrbook för de SEK I*. Hamburg: Quickborn-Verlag.

Jellinghaus, Hermann. 1892. *Die niederländischen Volksmundarten nach den Aufzeichnungen der Niederländer*. Norden, Leipzig: Diedr. Soltau's Verlag.

Kahl, Heinrich & Heinrich Thies. 2009. *Der neue SASS: Plattdeutsches Wörterbuch*. Neumünster: Wachholtz.

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

Lasch, Agathe. 1974. *Mittelniederdeutsche Grammatik* (Sammlung kurzer Grammatiken germanischer Dialekte 9). Halle (Saale): Max Niemeyer Verlag.

Leopold, Johan A. & Lubbertus Leopold. 1882. *Van de Schelde tot de Weichsel*. Groningen: J.B. Wolters.

Lindow, Wolfgang, Dieter Möhn, Hermann Niebaum, Dieter Stellmacher, Hans Taubken & Jan Wirrer. 1998. *Niederdeutsche Grammatik*. Leer: Schuster.

Lücht, Wilko. 2016. *Ostfriesische Grammatik*. Aurich: Ostfriesische Landschaftliche Verlags- und Vertriebsgesellschaft mbH.

Malmasi, Shervin & Marcos Zampieri. 2017. German dialect identification in interview transcriptions. In Preslav Nakov, Marcos Zampieri, Nikola Ljubešić, Jörg Tiedemann, Shevin Malmasi & Ahmed Ali (eds.), *Proceedings of the Fourth*

Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial), 164–169. Valencia: Association for Computational Linguistics. DOI: 10.18653/v1/W17-1220.

Nerbonne, John. 2009. Data-driven dialectology. *Language and Linguistics Compass* 3(1). 175–198.

Niebaum, Hermann. 2008. Het Nederduits. In Henk Bloemhoff, Jurjen van der Kooi, Hermann Niebaum & Siemon Reker (eds.), *Handboek Nedersaksische Taal- en Letterkunde*, 430–447. Assen: Koninklijke Van Gorcum.

Pedregosa, Fabian, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot & Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12. 2825–2830.

Pennington, Jeffrey, Richard Socher & Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. http://www.aclweb.org/anthology/D14-1162.

Peters, Friedrich Ernst. 1986. *Baasdörper Krönk*. Wolfgang Lindow & Paul Selk (eds.). Husum: Husum Druck- und Verlagsgesellschaft.

Pheiff, Jeffrey. 2022. Grammatikalisierung im Raum? Zu Variation und Wandel des Definitartikels in den niedersächsischen Dialekten Groningens und Drenthes. *Niederdeutsches Jahrbuch: Jahrbuch des Vereins für niederdeutsche Sprachforschung* 145. 130–155.

Popkema, Jan. 2018. *Grammatica Fries*. Leeuwarden: Afûk.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton & Christopher D. Manning. 2020. 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*. https://nlp.stanford.edu/pubs/qi2020stanza.pdf.

Saltveit, Laurits. 1983. Syntax. In Gerhard Cordes & Dieter Möhn (eds.), *Handbuch zur niederdeutschen Sprach- und Literaturwissenschaft*, 279–333. Berlin: Erich Schmidt Verlag.

Schallert, Oliver, Alexander Dröge & Jeffrey Pheiff. 2018. Doubly-filled COMPs in Dutch and German: A bottom-up approach. https://lingbuzz.net/lingbuzz/003979.

Schmid, Tanja. 2005. *Infinitival syntax: Infinitivus pro participio as a repair strategy* (Linguistik aktuell/Linguistics today 79). Amsterdam: John Benjamins.

Schröder, Ingrid. 2004. Niederdeutsch in der Gegenwart: Sprachgebiet – Grammatisches – Binnendifferenzierung. In Dieter Stellmacher (ed.), *Niederdeutsche Sprache und Literatur der Gegenwart*, 35–97. Hildesheim, Zürich, New York: Georg Olms Verlag.

Siewert, Janine, Yves Scherrer & Martijn Wieling. 2022. Low Saxon dialect distances at the orthographic and syntactic level. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, 119–124. Dublin, Ireland: Association for Computational Linguistics. DOI: 10.18653/v1/2022.lchange-1.12.

Siewert, Janine, Yves Scherrer, Martijn Wieling & Jörg Tiedemann. 2020. LSDC: A comprehensive dataset for Low Saxon dialect classification. In Marcos Zampieri, Preslav Nakov, Nikola Ljubešić, Jörg Tiedemann & Yves Scherrer (eds.), *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, 25–35. Barcelona (Online): International Committee on Computational Linguistics (ICCL). https://www.aclweb.org/anthology/2020.vardial-1.3.

Smits, Tom. 2011. Dialectverlies en dialectnivellering in Nederlands-Duitse grensdialecten. *Taal en Tongval* 63(1). 175–196.

Spruit, Marco René. 2008. *Quantitative perspectives on syntactic variation in Dutch dialects*. Utrecht: LOT.

Stellmacher, Dieter. 1983. Neuniederdeutsche Grammatik: Phonologie und Morphologie. In Gerhard Cordes & Dieter Möhn (eds.), *Handbuch zur niederdeutschen Sprach- und Literaturwissenschaft*, 238–278. Berlin: Erich Schmidt Verlag.

Thies, Heinrich. 2010. *SASS Plattdeutsche Grammatik*. Neumünster: Wachholtz.

van Bree, Cor. 2008. Syntaxis. In Henk Bloemhoff, Jurjen van der Kooi, Hermann Niebaum & Siemon Reker (eds.), *Handboek Nedersaksische Taal- en Letterkunde*, 113–133. Assen: Koninklijke Van Gorcum.

van der Vliet, Goaitsen. 2021. *Twentse Taalbank*. http://www.twentsetaalbank.nl/. Accessed: (15 December, 2021).

Wieling, Martijn & John Nerbonne. 2015. Advances in dialectometry. *Annual Review of Linguistics* 1(1). 243–264.

Wisser, Wilhelm. 1921. *Wat Grotmoder vertellt: Ostholsteinische Volksmärchen*. Jena: Eugen Diederichs.

Wolk, Christoph & Benedikt Szmrecsanyi. 2016. Top-down and bottom-up advances in corpus-based dialectometry. In Marie-Hélène Côté, Remco Knooihuizen & John Nerbonne (eds.), *The future of dialects: Selected papers from Methods in Dialectology XV*, 225–244. Berlin: Language Science Press.