LSDC - A comprehensive dataset for Low Saxon Dialect Classification

Siewert, Janine; Scherrer, Yves; Wieling, Martijn; Tiedemann, Jörg

Published in:
Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects

Publication date:
2020

LSDC – A comprehensive dataset for Low Saxon Dialect Classification

Janine Siewert
University of Helsinki
janine.siewert@helsinki.fi

Yves Scherrer
University of Helsinki
yves.scherrer@helsinki.fi

Martijn Wieling
University of Groningen
m.b.wieling@rug.nl

Jörg Tiedemann
University of Helsinki
jorg.tiedemann@helsinki.fi

Abstract

We present a new comprehensive dataset for the unstandardised West-Germanic language Low Saxon covering the last two centuries, the majority of modern dialects and various genres, which will be made openly available in connection with the final version of this paper. Since so far no such comprehensive dataset of contemporary Low Saxon exists, this provides a great contribution to NLP research on this language. We also test the use of this dataset for dialect classification by training a few baseline models comparing statistical and neural approaches. The performance of these models shows that in spite of an imbalance in the amount of data per dialect, enough features can be learned for a relatively high classification accuracy.

1 Introduction

Compared with the dominant languages of larger countries, minority languages tend to be underrepresented in terms of access to NLP tools. Availability of such tools however is of vital importance, since a lack of these indirectly forces groups already under pressure of language shift to resort to tools in the dominant language, with the consequence of a further decrease in the proportion of domains where the language can be used in daily life (Kornai, 2013). This is especially true for unstandardised languages like Low Saxon, where the lack of a written norm poses challenges for the development of modern NLP applications, which typically rely on large amounts of, ideally, orthographically uniform data. While a reference corpus exists for Middle Low Saxon (ReN-Team, 2019), the few datasets of modern Low Saxon available so far tend to either be very restricted content-wise (e.g. the DSA data (Wrede et al., 1927–1956)) or only represent a fraction of the language area without indication of the dialect (e.g. the OPUS data (Tiedemann, 2012)). In addition, both the DSA data and most of the OPUS data consist of content translated into Low Saxon instead of original texts, which will affect the naturalness of the language. The aim of this dataset for Low Saxon is thus to provide open and for the most part original data in Low Saxon covering nearly the whole language area in order to foster research and facilitate the development of NLP tools.

The composition of this dataset and testing the suitability of language recognition tools is a first step in our larger research project on processing Low Saxon data and modelling the historical development of the language-internal variation. Successful dialect recognition could thus be a useful step in a preprocessing pipeline where it would then be followed by normalisation to one of the writing systems in use, before applying tools developed for standardised languages to Low Saxon text.

In this paper, we will first give an overview of the societal and historical background as well as characteristic features of Low Saxon dialects, followed by a description of the dataset.[1] We will conclude with the presentation of a few baseline models for dialect identification trained on this data and an analysis of their results.

[1]The dataset is made available under a CC NC-BY-SA licence at https://github.com/Helsinki-NLP/LSDC/.

2 Background

Low Saxon is an unstandardised West-Germanic language with most of its around 5 million speakers today living in the north of Germany and the north-eastern parts of the Netherlands (Moseley, 2010). Starting from the 20th century, the usage of Low Saxon and its intergenerational transmission have been in decline but in the late 20th and in the 21st century, minor revitalisation and popularisation efforts have emerged, e.g. by introducing Low Saxon as a school subject and encouraging young musicians to produce songs in the language. Furthermore, especially since the advent of social media, more people have started to use Low Saxon as a written language for communication in daily life, as described e.g. by Palmiotta (2019) in his PhD thesis on Low Saxon speaking communities on Facebook.

Even though Low Saxon has some official status in several federal states of Germany, e.g. in Schleswig-Holstein (Landesregierung Schleswig-Holstein, 1992), and is protected under the European Charter for Regional and Minority Languages in both Germany and the Netherlands (Council of Europe, 2020), there is no official standard variety in use. This is why characteristics of local dialects tend to be rather well reflected in modern written Low Saxon, not only in terms of lexicon and syntax, but also in the writing system employed. These local writing traditions are based to different degrees and in different ways on the majority language orthography, i.e. German or Dutch. Some of their characteristics will be explained in more detail below.

2.1 Historical background an modern Low Saxon dialects

During the late Middle Ages, Low Saxon, as the main language spoken by Hanseatic merchants, played a major role as a language of international trade in Central and Northern Europe. However, in spite of the interregional influence of the writing tradition of Lübeck, the capital of the Hanseatic League, the Middle Low Saxon written language never became fully uniform (Stellmacher, 1990), and when the Hanseatic League lost its influence in the 16th and 17th century, the literary language started to fade out of use and was gradually replaced by Dutch and German in most written domains (Gabrielson, 1983).

While occasional Low Saxon texts were produced in the 17th and 18th century, the renaissance of Low Saxon as a written language is generally considered to have taken place in the 19th century, led by authors like Klaus Groth and Fritz Reuter, both represented in the LSDC (”Low Saxon Dialect Classification”) corpus.

Low Saxon today exists in the form of a dialect continuum without a common overarching written form. There are various ways of classifying these dialects based for instance on historical or current political units or certain isoglosses such as the usage of specific inflectional suffixes or particular vowel mergers.

Our classification of the Dutch Low Saxon dialects follows the division used by Bloemhoff et al. (2008) and the dialects from the German side were divided according to the traditional classification presented e.g. by Schröder (2004) and Stellmacher (1983). Table 2 and Figure 1 show which dialects are spoken in which country. This traditional classification is based on certain developments in the phoneme system and particular morphological features, and is still widely in use and often cited in standard works (Schröder, 2004, 51–52). As data for all subdialects from Germany was not easily obtainable or identifiable, in several cases, only the larger dialect group was used as category. For instance, whereas for Westphalia, all of the data could clearly be identified as belonging to either of three out of the four subdialects (MON, OWL and SUD), this was not the case for most of the other regions.

2.2 Phonological and morphological differences

Low Saxon is known for not having a differentiation into 1st, 2nd and 3rd person in the plural of verbs, but the dialects differ as to which suffix occurs. (Schröder, 2004, 43–44) A morphological feature commonly used in Low Saxon dialect classification thus is the plural suffix of verbs in the present tense. In eastern Low Saxon (MKB, MAR, NPR), East Frisian (OFR) and Gronings (GRO), the plural suffix is *-(e)n*, whereas the remaining dialects in this dataset use the suffix *-(e)t*. A characteristic morphological feature contrasting Eastphalian with the other dialects are the inflected forms of the personal pronouns, where e.g. *mik* ’me’ and *dik* ’you-SG (”thee”)’ are used instead of variants of *mi* and *di* elsewhere (Schröder,

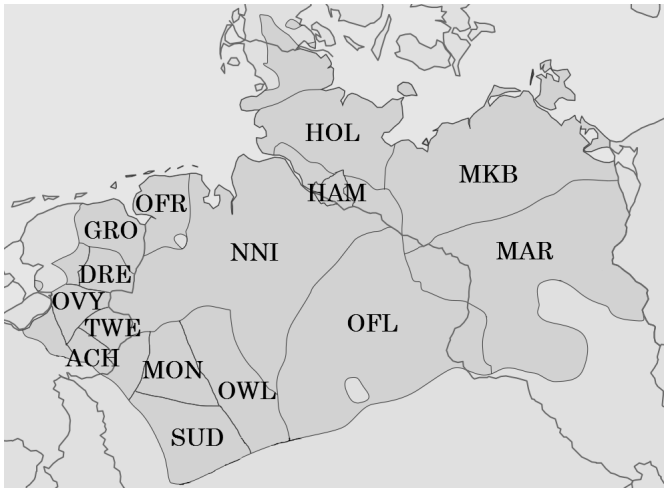

Figure 1: The geographic situation of the Low Saxon dialects covered in this article. The area of the Lower Prussian ('NPR') dialects, previously spoken further to the east along the coast of the Baltic Sea, is missing from the map. The dialect borders added (ACH, DRE, HAM, MON, OVY, OWL, SUD and TWE) are not meant to be precise delineations of the dialect areas, but are intended to give an impression of the position in relation to the other dialects. Map source: `https://commons.wikimedia.org/wiki/File:Low_Saxon_dialects.png`

2004, 49). In Dutch Low Saxon, a morphological change is attested in the dataset. In most Dutch Low Saxon dialects, the old second person singular *doe~du* and its corresponding verb inflection have fallen out of use today and have been replaced by the counterparts of Dutch *jij* 'you-SG' and *jullie* 'you-PL'. In texts from the 19th century, however, the old second person singular is still encountered.

On the phonological level, a feature usually employed for dialect classification is the development of stressed old short vowels in open syllable. These old short vowels in open syllable, e.g. Old Saxon *fugal* 'bird' and *etan* 'to eat' (Orel, 2003), were diphthongised in Middle Low Saxon time and either preserved as diphthongs or monophthongised (Lasch, 1914). Examples of these developments are the diphthongs in *Vuëgel* and *iäten* in the Münsterland dialect (Kahl, 2009) contrasting with *Vågel* and *äten* in Mecklenburg-Vorpommern (Herrmann-Winter, 2006). Subsequently, mergers occurred in several dialects. While the Westphalian dialects (MON, OWL and SUD) have preserved seven distinct reflexes of the lengthened/diphthongised short vowels, the northern dialects (HAM, HOL, MKB and NNI) only know a threefold distinction and Eastphalian (OFL) takes an intermediate position with five phonemes (Schröder, 2004, 53). According to map 8 in Panzer and Thümmel (1971), some of the central Dutch Low Saxon dialects have preserved a differentiation similar to the Westphalian dialects in Germany, while the remainder takes a more intermediate position similar to Eastphalian with 4–5 distinct phonemes.

2.3 Differences in writing systems

In addition to the above-mentioned differences in the dialects themselves, the influence of the majority language orthography introduces further divergences on the text level. The different grapheme usage in the German and the Dutch orthography thus causes the same pronunciation to appear clearly distinct in written form on the other side of the border: E.g. while according to the East Frisian online dictionary (Ostfriesische Landschaft, 2020), one should write *Huus* 'house', *för* 'for' and *südelk* 'southern', this corresponds to *hoes*, *veur* and *zudelk* on the other side of the border in Groningen (Reker, 2020). These divergent written forms do not represent different phonemes, but are a result of different graphemes used to represent the same phoneme following the usage in the Dutch and the German orthography.

Furthermore, there are also differences in the way the majority language orthography functions as a reference. This is expecially noticeable in the German Low Saxon writing systems. While the writing systems of the northern Low Saxon dialects in Germany use the written form of lexemes in the German orthography as a reference, the Westphalian writing systems more consistently adopt phoneme-grapheme

ACH:	Ziene olders hadden altied hard ewarkt en wazzen gezene leu in den naoberschop.
DE:	*Seine Eltern hatten immer hart gearbeitet und waren geschätzte Leute in der Nachbarschaft.*
NL:	*Zijn ouders hadden altijd hard gewerkt en waren voorname mensen in de gemeenschap.*
GRO:	Daor, kiek man ijs goud, 't kan best wezen, dat 't nog familie van die is.
DE:	*Da, guck nur mal gut, es kann gut sein, dass das noch Familie von dir ist.*
NL:	*Daar, kijk maar even goed, het kan best zijn dat 't nog familie van je is.*
HOL:	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.
DE:	*Arfest nahm zwei Karten für die erste Klasse und als ich darüber große Augen machte, lachte er und sagte, das könne darauf stehen, ich solle nur einsteigen.*
NL:	*Arfest nam twee kaarten voor de eerste klasse, en toen ik daarover grote ogen opzette, lachte hij en zei, het kan erop staan, maar ik zou gewoon instappen.*
MAR:	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.
DE:	*Und so, wie der Tod es dem Fischer erzählt hat, ist es auch gekommen; das ganze Dorf ist ausgestorben, aber der Fischer ist übriggeblieben und ist ein reicher Mann geworden, und seine Kinder leben noch bis auf diesen Tag in Göttin und sind reiche Leute.*
NL:	*En zo, hoe de dood het de visser verteld heeft, is het ook gebeurd; het hele dorp is uitgestorven, maar de visser is overgebleven en is een rijke man geworden, en z'n kinderen leven nog tot op de dag van vandaag in Göttin en zijn rijke mensen.*
OFL:	Ik kann nich sä güt wiet lupen un dorumme schölle mik miene Fründin hier ne Parkbuchte friehulen.
DE:	*Ich kann nicht so gut weit gehen und darum sollte mir meine Freundin hier eine Parkbucht freihalten.*
NL:	*Ik kan niet zo goed ver lopen en daarom moet mijn vriendin hier een parkeerplaats vrijhouden.*
SUD:	Eunige Dage später frogere de Magister, biu de veuer Johrestyien herren: Hiärmen sprank op, un de Magister mennte all, hai härr' et wieten.
DE:	*Einige Tage später fragte der Magister, wie die vier Jahreszeiten hießen: Harmen sprang auf und der Magister dachte schon, dass er es gewusst hätte.*
NL:	*Enige dagen later vroeg de magister, hoe de vier jaargetijden heetten: Harmen sprong op en de magister dacht al dat hij het had geweten.*

Table 1: Example sentences from six dialects and added translations into German (DE) and Dutch (NL).

correspondences from the German orthography: E.g. the SASS writing system (Kahl and Thies, 2009), nowadays used for part of the north-western dialects, and the writing system for Mecklenburg-Vorpommern (Herrmann-Winter, 2006) prescribe the usage of <h> as a vowel-length marker, if the same is used in the German cognate, as in *föhlen* and *fäuhlen* 'to feel' following German *fühlen*. Similarly, in the same writing systems, the Low Saxon phoneme /d/ is represented by the grapheme <t> or <tt> word finally, if this grapheme occurs in the German cognate. Examples of this are the words *Brett* 'board' and *wiet* 'wide', corresponding to German *Brett* and *weit*. In contrast, in the Münsterland writing system (Kahl, 2009), one would write *fölen*, *Bräd* and *wied*, according to the Low Saxon phonemes instead. However, most texts from the northern dialects included in the LSDC corpus predate these writing systems, so the authors did not necessarily adhere to the same rules. As a consequence of these differences in spelling, the LSDC dataset is not suitable for measuring distances between dialects without introducing a normalisation step. In order to illustrate the dialectal and orthographical variation, a selection of example sentences from LSDC is shown in Table 1.

3 The LSDC dataset

We have gathered a comprehensive data set of Low Saxon dialects covering nearly the whole language area. The collection includes historical texts from the 19th and 20th century as well as contemporary Low Saxon and therefore spans a period of around 200 years with most dialects being presented in at least two centuries. The total size is 105 876 sentences with an average sentence length of 20.16 words and a total word count of 2 134 753. This LSDC dataset presents a unique resource for modern Low Saxon which will be made openly available together with the final version of the paper. In the following subsections, we will provide background on the original text sources and the characteristics of the data.

3.1 Text sources

Most of the data for German Low Saxon dialects is copyright-free material from Wikisource. This is true for the dialects of Holstein, Hamburg, Mecklenburg-Vorpommern, Mark-Brandenburg, Lower Prussia and the north of Lower Saxony. The data for the Sauerland dialect originates from the Christine Koch Mundartarchiv and the Eastern Westphalian data consists partially of works written by Heinrich Stolte and made publicly available online by Olaf Bordasch, partially of texts from the website Lippisch Platt. Eastphalian and East Frisian data was provided by local authors and for the Münsterland dialects, we were given permission to use the data from Dr. Klaus-Werner Kahl's website.

The Groningen data originates from the online magazine Kreuze. Except for several older texts found on Wikisource, the remainder of data in Dutch Low Saxon dialects (Achterhoeks, Drents, Western Overijssels and Twents) was directly sent to us by different Low Saxon institutions or authors.

The links to the websites where the publicly accessible data can be found are listed in Table 5.

3.2 Description of the data

While a reference corpus for Middle Low Saxon (spanning the period 1200–1650 and excluding the dialects from today's Netherlands) exists (ReN-Team, 2019), so far no balanced corpus of more modern Low Saxon is available, which is why we hope that the LSDC dataset will greatly improve the possibilities for Low Saxon NLP research. The dataset spans the whole period from the renaissance of Low Saxon as a written language in the 19th century until today, covers nearly the whole of the current language area, and in addition one (nearly) extinct dialect, and presents the language in various genres. With the exception of passages from the Bible and other occasional translated works, the majority of texts was originally written in Low Saxon and therefore provides an authentic picture of the kind of language used in contemporary Low Saxon literature.

The most common genre in the dataset are short stories and short novels, and – especially in texts from the 19th and early 20th century – also fairytales and legends. In addition, the corpus contains genres as varied as religious texts, historical accounts, journal articles, poetry and songs, political speeches and simple texts for school children. An overview of the genres represented in the different dialects can be found in Table 2. Due to time restrictions, we however could not annotate the genres yet to make use of them in the experiments described in section 4.

The original text sources were not readily available in plain text format, but needed to be converted, and in particular the PDF documents required manual correction. Sentence splitting was performed using Python's NLTK tokenize package[2]. Moreover, we removed larger passages in other languages and performed careful manual cleaning of the test set.

3.3 Placenames

In addition to actual text data, we also collected Low Saxon place names from the two Low Saxon Wikipedia versions[3] and for the German side also from the websites of the district of Lüneburg and the local organisation Fehrs-Gill, also to be found in Table 5. These placenames will be included in the Low Saxon dataset to be published in connection with the article. The idea behind this setting is to test whether placenames could play a major role in dialect classification for Low Saxon, that we will discuss in more detail in the next section below. For the Lower Prussian region, we only found less than 30 placenames, so we excluded this dialect from the purely placename-based tests. Issues concerning the placename-based testing are on the one hand the incomplete lists on Wikipedia and furthermore the fact that the lack of standardisation of course also applies to the written form of placenames, so they do not necessarily occur in the same form in the lists and in the actual text data. Moreover, in the eastern regions of the Low Saxon language area, many placenames are of Slavonic origin, reflecting their history of settlement, and hence do not correspond to regular Low Saxon words.

[2] `https://www.nltk.org/`

[3] `https://nds-nl.wikipedia.org/wiki/` for Dutch Low Saxon and `https://nds.wikipedia.org/wiki/` for German Low Saxon.

Dialect region	Abbr.	Country	Sentences in train and test set		Tokens	Types	Centuries covered	Genres covered
Achterhoek	ACH	NL	500	488	20253	4020	20th, 21st	N
Drenthe	DRE	NL	5659	1000	97311	10538	19th, 21st	N
Groningen	GRO	NL	15999	1000	260477	28654	20th, 21st	various
Hamburg	HAM	DE	6095	1000	100590	9740	19th, 20th	N, T
Holstein	HOL	DE	11818	1000	263084	19554	19th, 20th	F, N
Mark-Brandenburg	MAR	DE	100	72	7102	1959	19th	F
Mecklenburg-Vorpommern	MKB	DE	14432	1000	541273	33122	19th	F, N
Münsterland	MON	DE	400	361	13468	3482	20th, 21st	N, S
Northern Lower Saxony	NNI	DE	401	377	18812	3592	20th, 21st	F, A
Lower Prussia	NPR	DE	200	155	9059	2751	19th, 20th	F, N, S
Eastphalia	OFL	DE	8377	1000	176187	15074	19th, 20th, 21st	various
East Frisia	OFR	DE	150	90	4051	1158	19th, 21st	N, A, S
Overijssel (west)	OVY	NL	800	547	21397	3267	21st	N, S
Eastern Westphalia	OWL	DE	14131	1000	260612	16319	20th, 21st	various
Sauerland	SUD	DE	16056	1000	329920	38831	19th, 20th, 21st	various
Twente	TWE	NL	368	300	11562	3231	21st	various

Table 2: Dialects and training set size. Abbreviations used: A = administration, announcements and politics, F = fairytales and legends, N = (short) stories and (short) novels, S = songs and poetry, T = theatre plays. Subcorpora marked with 'various' cover more than three genres, in addition to the ones mentioned above including e.g. religious texts, meta-discussions about language (usually about Low Saxon) and discussions of history.

4 Dialect identification

So far, Low Saxon is still an underresearched language within the field of NLP, which probably at least partly is due to its lack of a standardised form. We are not aware of any comparable previous work on dialect identification for this language, but Birkenes (2018) conducted n-gram-based measurements of the distance between dialects of German Low Saxon for dialect classification and identification of dialect areas using the Wenker atlas data (Wrede et al., 1927–1956).

For testing the usability of the dataset, we chose the fastText and langID toolkits to train baseline models. Both are supervised classifiers where the classes are predefined, but otherwise they differ in both the way the language data is represented and their general architecture.

4.1 Approaches to automatic language identification

There is a vast amount of literature on language identification and most approaches base their predictions on character n-gram statistics and language model features such as estimated token probabilities. A comprehensive overview of approaches is provided by Jauhiainen et al. (2018). There are also recent approaches based on neural models and representation learning. In our work we focus on two popular tools representing purely statistical models (langID, Lui and Baldwin (2011)) and neural models (fastText, Joulin et al. (2016)).

While fastText is also used for language identification, it is designed as a general text classification model, which is used for tasks like sentiment analysis as well. This orientation towards general text classification presumably is the reason for choosing a language representation based on bag of words and bag of word n-grams. If Low Saxon had a unified orthography, one would expect such a model to be more useful for dialect identification, which in this case would have to rely more on differences in lexicon and syntax.

The basic structure is a linear classifier which is combined with a rank constraint supposed to improve the generalisation of the model in case that some classes only have a small amount of examples. They use stochastic gradient descent, a linearly decreasing learning rate and hierarchical softmax in order to reduce training time. The best performance was achieved with word n-grams up to 5, but for our experiments we kept the default of 1.

The langID model is specifically designed for language identification controlling for divergent language use in different genres by choosing features with a high information gain related to language, but

a low information gain in relation to domain. Since our data however is not (yet) divided according to domain, a possibly useful functionality could thus not be taken advantage of.

In addition, a bias towards more high-resource varieties is supposed to be prevented by choosing a fixed number of features per variety. Unlike the fastText model, the features selected by langID for creating a document vector are not complete words, but character n-grams (1 to 4 grams), and no assumption is made concerning word delimitation. Due to the lack of a standard orthography, the same word may easily appear in slightly divergent spellings within the same dialects, which is why a character-based model seems more appropriate. However, Jauhiainen et al. (2018, 17) had determined that a larger unit, namely syllables and syllable n-grams are particularly suitable if the varieties to be identified are closely related. For classification, langID applies a multinomial naïve Bayes model with feature selection based on an information gain measure.

4.2 Discriminating Low Saxon from other Germanic languages

We originally chose the fasttext model because it includes a pretrained language model for Low Saxon ('nds'). This pretrained model however performed poorly on the majority of dialects, which were misclassified most frequently as Dutch or German, in all likelihood depending on whether they were written with a Dutch- or a German-based spelling. An exception are the north-western dialects from the German side, with a classification accuracy of more than 50% achieved for the dialects from East Frisia, Northern Lower Saxony, Hamburg and Schleswig-Holstein. This is unsurprising given that apparently the language model was trained on data from the German Low Saxon Wikipedia. Therefore, we eventually decided to train new Low Saxon models completely from scratch in order to not propagate the apparent north-western bias towards the dialect models.

4.3 Identifying Low Saxon dialects

The division into train and test set was based on the amount of dialect data. While the test set of more high-resource dialects was given a fixed size of 1000 sentences with the remainder being used for training, the test set size for the dialects with less than 1000 sentences of overall data was set to be at roughly 1-1.5:2 compared with the train set. The ratio however changed a little after removing duplicates from both the train and the test set. Furthermore, sentences of length three words or shorter, as well as sentences mostly or fully in a language other than Low Saxon were moved from the test set to the train set and replaced by fully Low Saxon sentences taken from the train set. The final size of the train and test sets per dialect is shown in Table 2.

During testing, the models were to classify the sentences separately instead of as a whole document. We considered this a meaningful size, since Low Saxon data for future research could be collected from social media platforms such as Facebook, Instagram or WhatsApp, where messages often do not exceed this length either.

Sampled training data

Since the fastText model exhibited a strong bias in favour of more resource-rich dialects, we decided to balance out the amount of training data. Thus, the train sets were either over- or undersampled to 10,000 sentences for each dialect. For oversampling, the same train sentences were copied over until a number of 10,000 or higher was reached. Subsequently, the train sets with a number higher than 10,000 were shuffled and cut off at 10,000 sentences. This was repeated three times to create three train sets per dialect, one for each of the three runs, as to not overly influence the final result by which part of the train set is dropped.

Placenames

In order to control for the effect of placenames on classification accuracy, we built a simple classifier where the dialect identification only depended on whether a placename from the list occurs in the sentence to be classified. This classifier's overall accuracy remained below 1%, which is unsurprising given the spelling variation and that most sentences do not contain any placenames to begin with. Therefore, placenames alone probably do not play a major role for dialect classification in our experiments.

	fastText		langID		placenames
	recall	precision	recall	precision	in train set
ACH	$0.0^{\pm 0.0}$	$0.0^{\pm 0.0}$	$0.0^{\pm 0.0}$	$0.0^{\pm 0.0}$	196
DRE	$4.0^{\pm 2.5}$	$10.0^{\pm 2.3}$	$3.4^{\pm 0.6}$	$15.9^{\pm 1.5}$	527
GRO	$23.3^{\pm 10.0}$	$66.6^{\pm 16.7}$	$12.7^{\pm 3.0}$	$24.5^{\pm 2.9}$	505
HAM	$0.1^{\pm 0.1}$	$33.3^{\pm 47.1}$	$0.0^{\pm 0.0}$	$0.0^{\pm 0.0}$	113
HOL	$7.6^{\pm 3.0}$	$13.1^{\pm 0.7}$	$8.3^{\pm 2.5}$	$20.0^{\pm 3.6}$	2288
MAR	$23.1^{\pm 17.1}$	$2.1^{\pm 0.2}$	$2.2^{\pm 1.1}$	$1.1^{\pm 0.6}$	494
MKB	$0.0^{\pm 0.0}$	$0.0^{\pm 0.0}$	$0.0^{\pm 0.0}$	$0.0^{\pm 0.0}$	157
MON	$0.5^{\pm 0.7}$	$0.4^{\pm 0.5}$	$0.0^{\pm 0.0}$	$0.0^{\pm 0.0}$	328
NNI	$22.9^{\pm 4.0}$	$3.3^{\pm 0.1}$	$69.8^{\pm 3.9}$	$4.4^{\pm 0.0}$	4564
OFL	$38.6^{\pm 1.3}$	$9.4^{\pm 0.0}$	$19.0^{\pm 3.6}$	$6.7^{\pm 0.7}$	2804
OFR	$0.7^{\pm 1.0}$	$1.0^{\pm 1.4}$	$0.0^{\pm 0.0}$	$0.0^{\pm 0.0}$	379
OVY	$10.6^{\pm 10.2}$	$7.3^{\pm 0.1}$	$1.3^{\pm 0.5}$	$8.1^{\pm 1.3}$	334
OWL	$2.7^{\pm 3.9}$	$2.6^{\pm 3.6}$	$1.6^{\pm 0.8}$	$26.7^{\pm 4.6}$	652
SUD	$0.0^{\pm 0.0}$	$0.0^{\pm 0.0}$	$0.2^{\pm 0.1}$	$0.5^{\pm 0.6}$	452
TWE	$0.0^{\pm 0.0}$	$0.0^{\pm 0.0}$	$0.7^{\pm 0.7}$	$9.4^{\pm 10.3}$	196
Macro F-score	9.4		7.9		

Table 3: Results of the placename-based models.

Subsequently, we used the placename list to train fastText and langID models in order to test if enough information on e.g. common letter combinations for successful classification can be extracted from placenames alone (see Table 3). An interesting observation here is the noticeable bias towards the more high-resource dialects in both the fastText and the langID models, while langID results appear to be more stable over several training runs. These classifiers, too, only attain a low F-score, so it seems that one can safely conclude that the placenames alone do not provide a major contribution for Low Saxon dialect identification. For more detailed results cf. Table 3.

Results of the dialect identification models

In the discussion of the results, we will focus on three aspects: the performance on low-resource vs. more high-resource data, the overall correctness, and which dialects tend to be confused.

When examining the performance of the first fastText model which was trained on the unsampled data (cf. the 'basic' columns in Table 4), a striking difference can be observed between the low-resource and more high-resource dialects: The model only attained a mean recall of between 0 and 6.2% on dialects with less than 500 sentences training data, it stayed below 50% for dialects with less than 1000 sentences and achieved decent results of over 90% for most dialects with over 5000 train sentences. The only exception are the neighbouring dialects of Hamburg and Holstein which relatively often are confused with each other. At the same time, the precision remains low, at 59.8–70.9%, for most of the high-resource dialects, since data in the low-resource dialects tends to be misclassified as one of the more resource-rich geographically close dialects. This is well reflected by the low macro F-scores of the fastText models on the original train set, as can be seen in the last row of Table 4.

With sampled training data, the fastText model shows a clear improvement in the macro F-score. As is evident from the 'sampled' columns in Table 4, even though the low-resource dialects' recall still falls behind the other dialects, a noticeable gain compared with the models trained on unsampled data can be observed. In the precision of dialect detection, no apparent difference related to the amount of distinct training data can be seen, and with the exception of the Holstein dialect, which again often was confused with the Hamburg dialect, a score between 82.2 and 98.0% is attained. The langID model is more stable across different training runs, which is apparent both from the macro F-scores and from the lower standard deviation scores. Interestingly, the sampling of the train set does not lead to noticeable improvements in the langID models. Clearly, the langID models manage to extract the features just as well without copied data, but nevertheless, do not attain the accuracy of the sampled fastText models.

Generally, it can be observed that it is possible for the models, for the sampled fastText one even more so than for the langID models, to learn to distinguish between the different Low Saxon dialects relatively well. Furthermore, dialects are rarely confused across the Dutch-German border. This is expected, as the different writing traditions increase the grapheme-level distance between otherwise close dialects.

	fastText				langID			
	basic		sampled		basic		sampled	
	recall	precision	recall	precision	recall	precision	recall	precision
ACH	$28.2^{\pm 15.4}$	$56.3^{\pm 6.3}$	$77.0^{\pm 5.6}$	$89.3^{\pm 4.0}$	$73.8^{\pm 4.7}$	$78.7^{\pm 1.8}$	$80.9^{\pm 3.1}$	$75.0^{\pm 0.7}$
DRE	$93.1^{\pm 6.4}$	$64.1^{\pm 7.6}$	$92.5^{\pm 3.7}$	$90.1^{\pm 2.1}$	$85.3^{\pm 1.3}$	$78.4^{\pm 0.4}$	$83.4^{\pm 0.4}$	$83.8^{\pm 0.5}$
GRO	$97.0^{\pm 1.5}$	$73.3^{\pm 13.1}$	$95.8^{\pm 2.3}$	$89.4^{\pm 5.9}$	$89.4^{\pm 0.1}$	$83.5^{\pm 2.0}$	$89.5^{\pm 0.8}$	$86.9^{\pm 1.2}$
HAM	$59.8^{\pm 30.5}$	$82.9^{\pm 17.3}$	$77.3^{\pm 12.8}$	$86.3^{\pm 7.4}$	$73.6^{\pm 0.5}$	$75.2^{\pm 0.6}$	$70.9^{\pm 0.3}$	$77.0^{\pm 0.3}$
HOL	$87.1^{\pm 12.3}$	$66.8^{\pm 18.9}$	$95.4^{\pm 4.1}$	$63.2^{\pm 11.8}$	$80.8^{\pm 0.3}$	$71.7^{\pm 0.6}$	$77.9^{\pm 0.3}$	$73.0^{\pm 0.8}$
MAR	$0.0^{\pm 0.0}$	$0.0^{\pm 0.0}$	$63.9^{\pm 2.3}$	$97.3^{\pm 2.4}$	$60.2^{\pm 2.6}$	$62.8^{\pm 1.4}$	$66.7^{\pm 0.0}$	$41.3^{\pm 0.4}$
MKB	$96.9^{\pm 1.2}$	$72.0^{\pm 6.6}$	$80.3^{\pm 9.6}$	$95.7^{\pm 2.9}$	$86.0^{\pm 0.7}$	$77.0^{\pm 0.7}$	$84.2^{\pm 1.1}$	$80.2^{\pm 2.4}$
MON	$3.1^{\pm 2.3}$	$29.6^{\pm 40.9}$	$61.9^{\pm 7.7}$	$96.4^{\pm 3.4}$	$65.2^{\pm 1.7}$	$77.0^{\pm 2.1}$	$68.6^{\pm 1.0}$	$76.2^{\pm 0.7}$
NNI	$2.4^{\pm 2.7}$	$32.8^{\pm 29.7}$	$49.2^{\pm 7.8}$	$91.9^{\pm 4.9}$	$42.4^{\pm 0.4}$	$60.6^{\pm 2.0}$	$51.8^{\pm 0.6}$	$54.8^{\pm 1.3}$
NPR	$6.2^{\pm 0.6}$	$80.6^{\pm 14.6}$	$74.2^{\pm 1.1}$	$95.0^{\pm 0.6}$	$77.4^{\pm 1.4}$	$68.1^{\pm 1.3}$	$80.4^{\pm 1.1}$	$61.0^{\pm 0.7}$
OFL	$90.9^{\pm 6.6}$	$94.4^{\pm 3.2}$	$95.5^{\pm 1.3}$	$91.9^{\pm 1.6}$	$81.8^{\pm 0.1}$	$86.6^{\pm 0.7}$	$80.5^{\pm 0.6}$	$86.9^{\pm 0.7}$
OFR	$1.1^{\pm 0.9}$	$11.8^{\pm 8.5}$	$54.4^{\pm 5.5}$	$83.9^{\pm 9.0}$	$39.6^{\pm 2.1}$	$64.2^{\pm 4.0}$	$45.2^{\pm 1.9}$	$39.4^{\pm 1.7}$
OVY	$45.8^{\pm 25.4}$	$83.7^{\pm 2.8}$	$89.9^{\pm 1.6}$	$89.9^{\pm 2.8}$	$78.4^{\pm 2.4}$	$84.8^{\pm 3.0}$	$81.8^{\pm 1.3}$	$87.9^{\pm 1.3}$
OWL	$97.7^{\pm 1.1}$	$91.9^{\pm 2.7}$	$96.3^{\pm 1.7}$	$98.0^{\pm 1.1}$	$88.0^{\pm 0.4}$	$92.4^{\pm 0.2}$	$88.2^{\pm 0.1}$	$93.4^{\pm 0.0}$
SUD	$97.4^{\pm 0.8}$	$78.6^{\pm 5.8}$	$95.0^{\pm 1.6}$	$90.0^{\pm 3.5}$	$84.9^{\pm 0.9}$	$82.0^{\pm 0.6}$	$86.1^{\pm 0.8}$	$85.4^{\pm 0.7}$
TWE	$0.6^{\pm 0.4}$	$19.4^{\pm 14.2}$	$69.9^{\pm 4.8}$	$87.1^{\pm 4.0}$	$59.9^{\pm 1.1}$	$75.4^{\pm 2.5}$	$68.8^{\pm 2.6}$	$70.5^{\pm 1.8}$
macro-avg.	50.5	58.6	79.3	89.7	72.9	76.2	75.3	73.3
avg. F-scores	$55.0^{\pm 3.7}$		$84.2^{\pm 0.6}$		$74.5^{\pm 0.6}$		$74.3^{\pm 0.1}$	

Table 4: Mean and standard deviation for recall and precision of the three training runs of the different models and F-scores separately for the three training runs.

Furthermore, the confusion of dialects tends to correlate with geographical closeness. In the langID models, this is even more evident than in the fastText models, which seem to have a stronger bias towards particular dialects. E.g. the Münsterland dialect is most often misclassified as the neighbouring Sauerland dialect. We also observed that the low-resource Markish-Brandenburgish dialect is most often confused with the dialects from the northern neighbouring region Mecklenburg-Vorpommern. Also the fact that the most prominent confusion between high-resource dialects can be observed between Hamburg and Holstein suggests that the models do in fact learn relevant features and that some dialects simply are harder to classify correctly due to their greater similarity.

The performance of the models also correlates roughly with the intuition of the Low Saxon speaker in our group, who e.g. would judge the dialects from Holstein, Hamburg and Northern Lower Saxony present in the dataset to be comparatively similar and would have an easier time distinguishing the three Westphalian subdialects of Eastern Westphalia, the Sauerland and the Münsterland, even given that they speak a northern dialect themself. This may at least partly be due to the writing systems employed, since while there are distinct writing systems for at least Eastern Westphalia and the Münsterland, authors from Holstein, Hamburg and Northern Lower Saxony follow approximately the same tradition.

5 Conclusions, discussions and future work

The LSDC dataset presented in this paper is a comprehensive dataset for contemporary Low Saxon. With nearly all modern dialect groups being included and two centuries as well as various genres being covered, this dataset provides a unique new resource for NLP research on and application development for this minority language. We have tested the resource on a dialect identification task, demonstrating its use for further collections and classifications of dialectal data. Even though the amount of data per dialect is not balanced in LSDC, we have seen that sampling makes it possible to achieve a comparable precision for both high- and low-resource dialects with a neural classification model and that the performance of that model even approximately resembles native speaker intuition.

One central aspect affecting dialect identification are the various ways of writing Low Saxon. As no normalisation of spelling was included, we need to acknowledge that we cannot fully discern in how far the models classify the dialects based on the local writing system or based on actual dialect features such as lexicon, inflectional suffixes and syntax. As a consequence, it would be meaningful to conduct additional dialect identification experiments with orthographically normalised Low Saxon text data. In addition, one could test the effect of grouping the dialects according to similarity in writing systems and

training separate models for instance for Dutch Low Saxon, northern German Low Saxon and southern German Low Saxon.

The next steps in our research project on the historical development of the language internal variation in Low Saxon will require us to expand the corpus. First of all, more data in the low-resource dialects will be needed and the time coverage for all dialects needs to be improved, so that roughly the same amount of data is available for each dialect and century. Ideally, each dialect should additionally be represented by a variety of genres for each century and have the genre included as part of the annotation. This will allow us to investigate to what extent characteristics of a specific domain are misinterpreted as dialect features. Lastly, we will need to increase the time depth in order to close the gap between the Middle Low Saxon reference corpus and our modern Low Saxon dataset, so that the development of Middle Low Saxon into contemporary Low Saxon can be traced.

Acknowledgements

We would like to express our gratitude to the *Lännerzentrum för Nedderdüütsch* and the *Oostfreeske Landskupp* for establishing contact with Low Saxon authors, and especially to the authors themselves who provided us with data that would not have been accessible online: Diana Abbink from the *Erfgoedcentrum Achterhoek en Liemers*, Rolf Ahlers, Erich Bolinius, Marieke Dannenberg from the *Twentehoes*, Jan Germs from the *Huus van de Taol*, Joop Hekkelman, Dr. Harrie Scholtmeijer from the *IJsselacademie* and Gerard Uwland. Furthermore, we would like to thank the Discovery group lead by Prof. Hannu Toivonen at the department of Computer Science at the University of Helsinki for financing Janine Siewert's summer internship in 2019 in connection with which the collection of the dataset presented here was started.

References

Magnus Breder Birkenes. 2018. N-Gramm-basierte Ähnlichkeitsmessungen als dialektometrische Methode: Die Fragebögen des Wenker-Atlas. Linguistisches Kolloquium, University of Munich, Germany.

Henk Bloemhoff, Jurjen van der Kooi, Hermann Niebaum, and Siemon Reker, editors. 2008. *Handboek Nedersaksische Taal- en Letterkunde*. Koninklijke van Gorcum, Assen, Netherlands.

Council of Europe. 2020. Reservations and Declarations for Treaty No.148 - European Charter for Regional or Minority Languages.

Artur Gabrielson. 1983. Die Verdrängung der mittelniederdeutschen durch die neuhochdeutsche Schriftsprache. In Gerhard Cordes and Dieter Möhn, editors, *Handbuch zur niederdeutschen Sprach- und Literaturwissenschaft*, pages 119–153. Erich Schmidt Verlag, Berlin, Germany.

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

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2018. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65, 04.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Heinrich Kahl and Heinrich Thies. 2009. *der neue SASS – Plattdeutsches Wörterbuch*. Wachholtz Verlag, Neumünster, Germany.

Klaus-Werner Kahl. 2009. *Wörterbuch des Münsterländer Platt*. Aschendorff Verlag, Münster, Germany.

András Kornai. 2013. Digital language death. *PLOS ONE*, 8(10):1–11, 10.

Landesregierung Schleswig-Holstein. 1992. Allgemeines Verwaltungsgesetz für das Land Schleswig-Holstein (Landesverwaltungsgesetz - LVwG -) in der Fassung der Bekanntmachung vom 2. Juni 1992 § 82 b Regional- und Minderheitensprachen vor Behörden.

Agathe Lasch. 1914. *Mittelniederdeutsche Grammatik*. Max Niemeyer Verlag, Tübingen, Germany. Unchanged reprint from 1974.

Marco Lui and Timothy Baldwin. 2011. Cross-domain feature selection for language identification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 553–561, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

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

Vladimir Orel. 2003. *A Handbook of Germanic Etymology*. Brill, Leiden and Boston.

Ostfriesische Landschaft. 2020. Plattdeutsch-Hochdeutsches Wörterbuch für Ostfriesland. https://www.platt-wb.de/.

Michele Palmiotta. 2019. *Social network e lingue minoritarie: il gruppo Facebook come spazio di discussione e costruzione dell'identità linguistica del Niederdeutsch*. Ph.D. thesis, Università degli Studi di Bari Aldo Modo.

Baldur Panzer and Wolf Thümmel. 1971. *Die Einteilung der niederdeutschen Mundarten auf Grund der strukturellen Entwicklung des Vokalismus*. Max Hueber Verlag, München, Germany.

Siemon Reker. 2020. Groninger zakwoordenboek. http://www.groningsonline.nl/woordenboek.

ReN-Team. 2019. Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200-1650). Archived in Hamburger Zentrum für Sprachkorpora. Version 1.0. Publication date 2019-08-14.

Ingrid Schröder. 2004. Niederdeutsch in der Gegenwart - Sprachgebiet – Grammatisches – Binnendifferenzierung. In Dieter Stellmacher, editor, *Niederdeutsche Sprache und Literatur der Gegenwart*, pages 35–97. Georg Olms Verlag, Hildesheim and Zürich and New York.

Dieter Stellmacher. 1983. Neuniederdeutsche Grammatik – Phonologie und Morphologie. In Gerhard Cordes and Dieter Möhn, editors, *Handbuch zur niederdeutschen Sprach- und Literaturwissenschaft*, pages 238–278. Erich Schmidt Verlag, Berlin, Germany.

Dieter Stellmacher. 1990. *Niederdeutsche Sprache – Eine Einführung*. Peter Lang, Bern, Frankfurt am Main, New York and Paris.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*. Data available at: http://opus.nlpl.eu/.

Ferdinand Wrede, Walther Mitzka, and Berhard Martin, editors. 1927–1956. *Deutscher Sprachatlas auf Grund des Sprachatlas des Deutschen Reiches von Georg Wenker. Begonnen von Ferdinand Wrede, fortgesetzt von Walther Mitzka und Bernhard Martin*. N.G. Elwert'sche Verlagsbuchhandlung, Marburg (Lahn), Germany.

Dialect	Web address
GRO	http://kreuzekeuze.nl/kreuzewebstee/index.html
HOL placenames	https://sass-platt.de/plattdeutsche-ortsnamen-schleswig-holstein/index.html
MON	https://www.plattdeutsch.net/pages/platt-lesen/gedichte-und-geschichten.php http://www.sauerlandmundart.de/daunlots.html
NNI placenames	https://www.landkreis-lueneburg.de/Home-Landkreis-Lueneburg/Bildung-Soziales-und-Gesundheit-Landkreis/Bildung-und-Kultur/Kultur/Plattdeutsch.aspx
OWL	http://www.plattdeutsch-niederdeutsch.net/der_bauernhof_um_1870/index.htm http://www.plattdeutsch-niederdeutsch.net/neues_testament/index.htm http://www.lippischplatt.de/
SUD	http://www.sauerlandmundart.de/daunlots.html

Table 5: Origin of the Low Saxon data from online sources other than Wikisource and Wikipedia.