Towards a balanced annotated Low Saxon dataset for diachronic investigation of dialectal variation

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

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

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

Abstract

We describe a new annotated dataset for Low Saxon with the intention to complement existing corpora. This corpus covers the period from the 15th to the 21st century and is annotated with PoS and morphosyntactic tags as well as century and region information. This dataset will be used for diachronic dialectometry, but can lend itself to other NLP tasks as well. The target size is around 2000 sentences per dialect and century and at the time of writing, 798 texts have been selected for inclusion in the corpus. They will be gradually added as the annnotation progresses.

1 Introduction

We present a dataset for Low Saxon,[1] a Germanic minority language spoken by roughly five million people in Northern Central Europe (Moseley, 2010). Despite its relatively large number of speakers, there are hardly any annotated corpora for this language, hampering corpus-based research into more modern varieties and causing a lack of well-functioning NLP tools.

The dataset is part of our research into the diachronic development of the internal variation in Low Saxon and builds upon the Reference Corpus Middle Low German/Low Rhenish (1200-1650) (ReN-Team, 2019) (henceforth *ReN*) and the LSDC dataset (Siewert et al., 2020) attempting to fill the gap between them. Therefore, it covers both historical and contemporary Low Saxon dialects from the Veluwe in the western corner of the language area to the Lower-Prussian dialects in the east.

Our ultimate goal with this new dataset is to perform analyses of the internal variation within Low Saxon and its change over time. Questions of interest are, for instance, the frequency and geographical spread of features like two-part conjunctions (*dãrümme dat*, *êr dat*), double negation, and whether the perfect tense of modal verbs requires the main verb to occur in the infinitive or the perfect participle. These relate to the larger topic of interdialectal contact and how stable syntactic structures are when the speaker community is under constant exposure to a closely related more prestigious language and under pressure of language shift.

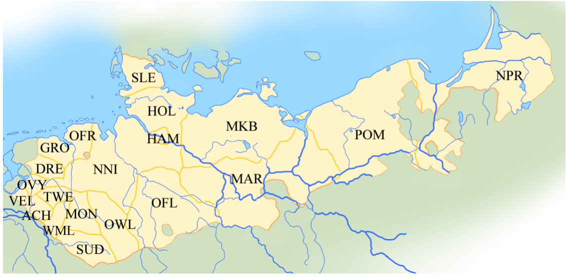

Figure 1: The Low Saxon dialects to be covered in the corpus.

2 Background

Low Saxon belongs to the western branch of the Germanic languages and is traditionally spoken mainly in Northern Germany and the North-Eastern Netherlands with official recognition in both countries. The eastern dialects Pomeranian (POM) and Low Prussian (NPR) shown in Figure 1 were spoken in these regions prior to WW II.[2]

When the Hanseatic League lost its status in the 16th and 17th century, the Middle Low Saxon literary language was replaced by southern varieties.

[1] Also called *Low German*, referring here to the varieties protected under the European Charter for Regional and Minority Languages as *Nedersaksisch* in the Netherlands and *Niederdeutsch* in Germany as well as extinct eastern varieties.

[2] The Baltic dialects previously spoken north of the Low Prussian area and included in the ReN will not form part of our corpus, as although e.g. in Estonia, Low Saxon had survived as a spoken language until the 19th, probably even until the early 20th century, (Ariste, 1981, 97–98) the amount of written post Middle Low Saxon sources preserved seems too small for meaningful analyses.

While Low Saxon survived in oral communication and occasional Low Saxon texts continued to be produced, German and Dutch became the dominant written languages (Gabrielson, 1983).

Despite its official recognition and usage in e.g. the media and school education today, no interregional standard language has been introduced so far, resulting in a tendency for Low Saxon speakers to follow regional writing traditions or come up with systems of their own, both of which are often based on the majority language orthography of the respective country.

The creation of NLP tools for modern Low Saxon thus requires a large annotated corpus representing this dialectal and orthographic variation.

3 Resources available

The main resources already available are the ReN and the LSDC. The ReN comes with HiNTS tags[3], morphological annotation and lemmatisation for the major regions of the northern dialects (North Low Saxon, East Elbian, Baltic Low Saxon), Westphalian, Eastphalian including Elbe Eastphalian, and South Marchian (*Südmärkisch*) spanning the time from ca. 1200 to 1650 (Peters and Nagel, 2014). The 146 annotated texts contain around 1.4 million tokens and the 89 transcribed (i.e. not annotated) ones ca. 900,000 tokens.

The LSDC dataset contains ca. 2 million tokens in ca. 100,000 sentences representing 16 dialects from the 19th century onward (Siewert et al., 2020). It covers a different set of dialects than the ReN, is smaller and includes neither PoS or morphological tags, nor lemmatisation.

Limitations of these datasets are, for instance, that the ReN excludes the dialects from today's Netherlands and that the LSDC dataset is only annotated for century and dialect, but does not include morphosyntactic annotation. In addition, the LSDC dataset is not very balanced, meaning that not all dialects are equally well represented in all of the three centuries covered.

The ASnA (*Atlas spätmittelalterlicher Schreibsprachen des niederdeutschen Altlandes und angrenzender Gebiete*) is based on a large collection of transcriptions of Middle Low Saxon documents excluding the eastern language area but including varieties from today's Netherlands (Peters, 2017). However, this dataset is not publicly available.

In addition to these, there is a thus far unpublished dataset from the University of Groningen / Centrum Groninger Taal en Cultuur for the Gronings dialect used by de Vries et al. (2021), which contains around 50k tokens, PoS tags and lemmatisation to standard Dutch, and might serve as additional training data for our tagging task.

A few larger corpus collections, such as OPUS or the Wikipedia dumps, contain Low Saxon data as well, but since generally no information on the dialect is provided, we decided to exclude them.

4 Data collection and selection

We are striving to gather at least 2000 sentences per dialect and century. Preferably, these should represent a variety of writers, genres and different places within the dialect region. For a somewhat balanced representation, the size of the geographical regions should at least be roughly comparable. Whereas in the LSDC dataset, the Westphalian group was subdivided into several subdialects both on the German and the Dutch side, our intention is to treat German Westphalian as one group and Dutch Westphalian as another one. More detailed information on the origin of the texts, e.g. the birth place of the writer or the printing place, will nonetheless be provided if available.

For German Low Saxon, we primarily collect data from the period between the middle of the 17th and the early 19th century, since this time span is covered by neither the ReN nor the LSDC; however, for Dutch Low Saxon it has been necessary to start our data collection from the 15th century. The LSDC provides a sufficiently large amount of sentences for some dialects and centuries, but most dialects still require additional data.

As we ultimately plan to perform syntactic analyses, we prefer prose, but the lack of data for various dialects, particularly in the 17th and 18th century, might necessitate an inclusion of poetry. In that case, genres will be labelled as well.

We have started to compile our own set of older Dutch Low Saxon data where the Middle Low Saxon data from Groningen and Drenthe mostly originate from the Cartago website,[4] from Twente from the Twentse Taalbank.[5] In addition, we also gather digitised data from local archives.

The German Low Saxon data mainly consists

[3] *Historisches-Niederdeutsch-Tagset*, adapted for Middle Low Saxon based on the HiTS (Historical Tagset for German) and explained by Barteld et al. (2018)

[4] http://cartago.nl/nl/

[5] http://www.twentsetaalbank.nl/

of digitised data from German university libraries and Google Books. Our search largely relies on Hansen's literature catalogue (Hansen, 2021), which strives to list all German Low Saxon authors as well as all books and other media published in German Low Saxon from 1473 onward.

Table 1 shows the number of texts collected so far. These texts differ largely in size, the shortest ones consisting of only one or a few pages and the longest ones being complete books with several hundreds of pages.

The data selection for older Dutch Low Saxon is not always straightforward. Even medieval writings from this area often contained both eastern (= Low Saxon) and western (= Dutch) traits (Niebaum, 1997, 63), and in contrast with the switch to the clearly distinct German in the areas further east in the 16th and 17th century the written language in the Dutch Low Saxon regions gradually shifts towards the comparatively similar one used in the Western Netherlands (Kremer, 2008, 43). Consequently, the question arises which texts are still sufficiently Low Saxon and which ones should instead by classified as Dutch and excluded from the corpus. A possible solution could be to base this on orthographic criteria. On the other hand, for the regions in Germany, it is generally easy to determine if a text is written in Low Saxon or German.

	15th	16th	17th	18th	19th
GLS			39	88	194
DLS	197	206	5		69

Table 1: Number of texts per group (*GLS* 'German Low Saxon' and *DLS* 'Dutch Low Saxon') and century.

The first version of the dataset will contain 200 sentences with manually corrected PoS and morphological annotation representing four dialects (Eastphalian, Holsatian, Marchian/Brandenburgish and Mecklenburgish - West Pomeranian) of German Low Saxon with 50 sentences each. The Mecklenburgish - West Pomeranian data stems from the second half of the 17th century, the Marchian/Brandenburgish data from the 18th century and the Holsatian and Eastphalian data from the first half of the 19th century. We will continuously update the dataset and add more sentences as the annotation progresses.

5 Preprocessing and annotation

Text acquisition Many of the digitised texts from the 17th, 18th and 19th century are only available as scans, while 59 of them include raw OCR. We have begun manual corrections of the raw OCR for training specialised models with Transkribus[6].

Sentence splitting General sentence splitting tools tend to work well on modern Low Saxon texts, but this is not the case for Early Modern Low Saxon and even less so for medieval texts, since punctuation does not follow the modern conventions. In the ReN, sentence splitting was based on the occurrence of inflected words. As a result, the corpus consists in large parts of sentence fragments instead of more complex sentence structures. While this might be an appropriate solution for the context of the Reference Corpus, it does not suit our goal of diachronic comparison of syntactic structures. Furthermore, this might pose difficulties to tagging, as disambiguation would often make it necessary to look across sentence fragment boundaries.

Morphosyntactic tagging The ReN serves as the basis for automatising the annotation process. We have converted the PoS and morphological tags in the ReN to the UD standard with a replacement script followed by manual corrections, since the correspondences do not always match one-to-one. For instance, the ReN often shows no distinction between conjunction and subjunction, and in several cases different usages of the same lemma are given the same PoS tag, such as only ADV in case of *of* 'if, or' even though it can function as both an adverb and a conjunction. Furthermore, following the ReN annotation, we have added extra labels for marking strong and weak declension, which do not belong to UD's universal features.

The converted ReN data is then used for training a full morphological tagger to annotate both the remainder of the ReN and the Middle Low Saxon data from the Netherlands. In a preliminary experiment with a small manually corrected Dutch Low Saxon dataset, a BiLSTM tagger (Scherrer and Rabus, 2019) trained on ReN data achieved a morphological tagging accuracy of around 85%.

We will manually correct a few hundred sentences of the automatic annotation and use those for fine-tuning. This process will be repeated step-by-step with data from the following century until

[6] https://readcoop.eu/transkribus/?sc=Transkribus

```
# sent_id = NDS_010_HOL_1910_as-noch-de-trankrusel-brenn
# text_orig = Ja, wo is de Knieptang?
# text = Ja, wår is de knyptange?
1  Ja         ja         INTJ   _  _                          0  root   _  lemma[gml]=jâ^1|SpaceAfter=No
2  ,          ,          PUNCT  _  _                          3  punct  _  _
3  wår        wår        ADV    _  _                          1  conj   _  lemma[gml]=wôr(e)
4  is         weasen     AUX    _  Number=Sing|Person=3       3  cop    _  lemma[gml]=wēsen^2
5  de         de         DET    _  Gender=Fem|Number=Sing     6  det    _  lemma[gml]=dê^1
6  knyptange  knyptange  NOUN   _  Gender=Fem|Number=Sing     3  nsubj  _  lemma[gml]=knîptange|SpaceAfter=No
7  ?          ?          PUNCT  _  _                          3  punct  _  _
```

Figure 2: Example of the UD annotation with reduced morphological features.

the contemporary period.

UD annotation Aside from the basic corpus, we have also started to select sentences to be included in a separate dataset for Universal Dependencies[7]. Mostly, these sentences originate from public domain texts included in the LSDC dataset, but we make use of our additional resources as well. This UD dataset will cover the Modern Low Saxon period, contain roughly the same amount of sentences per dialect and, in addition to PoS and morphological tags[8], it will feature dependencies and lemmatisation.

Due to the lack of an interregional standard, there is not one single obvious choice for lemmatising a dataset for Low Saxon covering several centuries and regions. As a compromise, we have opted for double lemmatisation: The main lemma will be given in the *Nysassiske Skryvwyse* – an interregional spelling used by e.g. the Dutch Low Saxon Wikipedia which is based on a historically motivated abstract set of common phoneme distinctions instead of a particular local pronunciation[9] – while a second lemma will be provided in normalised Middle Low Saxon following Lasch et al. (1928 ff) if the word was already attested at that stage of the language, cf. the tenth column in Figure 2.

6 Challenges

A few morphological issues require further discussion in relation to the annotation, since we need to take into account that we annotate language change in process. We will illustrate two of these issues here.

Differing number of inflectional categories Mergers of inflectional categories have occurred in different dialects at different points in time and to a different extent. For example, the distinction between dative and accusative still present in Middle Low Saxon has been lost in most modern dialects (Lindow et al., 1998, 144). Since the corpus contains both varieties with and without this distinction, our approach is to annotate as if the distinction had been preserved in all of the dialects. When, however, the local variety clearly shows a different inflection, i.e. if an accusative-like form is used instead of the expected dative, the regional annotation will be given in the tenth column in the form *Case[regional]=Acc*.

Change in pronoun usage The data we have collected shows that while it is not uncommon to still encounter the old 2nd person singular *dû* in Dutch Low Saxon texts from the 19th century, this pronoun has faded out of use in most dialects at the latest by the 21st century. With the exception of Groningen, North Drenthe and parts of Twente and the Achterhoek, Dutch Low Saxon dialects today have usually lost the *dû* (Bloemhoff, 2008, 101–103) and instead adopted the Standard Dutch system for the 2nd person using the counterparts of Dutch *jij* and *jullie* for the singular and plural respectively.

Due to the fact that the original pronoun of the second person plural *gî* was (and partly still is) also used as a polite address, one cannot always tell from the context if *gî* as referring to a single person is to be interpreted as a politeness marker or whether it already has replaced the *dû* as the default 2nd person singular. In such unclear cases, we refrain from annotating for number or politeness. By default, the *gî* and its agreeing verbs will nevertheless be annotated as plural in the sixth column with divergent regional developments being marked in the tenth column as *Number[regional]=Sing*.

[7] https://raw.githubusercontent.com/UniversalDependencies/UD_Low_Saxon-LSDC/master/nds_lsdc-ud-test.conllu

[8] The morphological tags are still missing in the first version, but will be included in the second one.

[9] https://skryvwyse.eu/

7 Access

The dataset can be accessed via our Helsinki NLP GitHub page[10]. The first release is published under a CC BY-NC licence, but as more data is added, different parts of the dataset might be published with separate licences depending on the licences the original files were provided with.

8 Conclusion

Our new balanced dataset for Low Saxon will cover the whole Modern Low Saxon period as well as late Middle Low Saxon from the Dutch side, and include not only annotation for dialect and century, but also PoS and morphological tags.

This novel resource will thus facilitate investigations into dialectal variation across time and, in addition, offer new possibilities to the development of NLP tools for this low-resource language.

References

Paul Ariste. 1981. *Keelekontaktid*. Valgus, Tallinn, Estonia.

Fabian Barteld, Sarah Ihden, Katharina Dreessen, and Ingrid Schröder. 2018. HiNTS: A tagset for middle low German. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Henk Bloemhoff. 2008. Klank- en vormleer. In Henk Bloemhoff, Jurjen van der Kooi, Hermann Niebaum, and Siemon Reker, editors, *Handboek Nedersaksische Taal- en Letterkunde*, page 65–112. Van Gorcum, Assen, Netherlands.

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.

Peter Hansen. 2021. Die niederdeutsche Literatur.

Ludger Kremer. 2008. Geschiedenis van de Nedersaksische taalkunde. In Henk Bloemhoff, Jurjen van der Kooi, Hermann Niebaum, and Siemon Reker, editors, *Handboek Nedersaksische Taal- en Letterkunde*, page 23–51. Van Gorcum, Assen, Netherlands.

Agathe Lasch, Conrad Borchling, Gerhard Cordes, and Dieter Möhn. 1928 ff. *Mittelniederdeutsches Handwörterbuch*. Wachholz Verlag, Neumünster, Germany.

Wolfgang Lindow, Dieter Möhn, Hermann Niebaum, Dieter Stellmacher, Hans Taubken, and Jan Wirrer. 1998. *Niederdeutsche Grammatik*. Verlag Schuster, Leer, Germany.

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

Hermann Niebaum. 1997. Ostfriesisch-groningische Sprachbeziehungen in Geschichte und Gegenwart. In Volkers F. Faltings, Alastair G.H. Walker, and Ommo Wilts, editors, *Friesische Studien III*, page 49–82. Odense University Press.

Robert Peters. 2017. *Atlas spätmittelalterlicher Schreibsprachen des niederdeutschen Altlandes und angrenzender Gebiete (ASnA)*. De Gruyter, Berlin and Boston.

Robert Peters and Norbert Nagel. 2014. Das digitale 'Referenzkorpus Mittelniederdeutsch / Niederrheinisch (ReN)'. *Jahrbuch für Germanistische Sprachgeschichte*, 5(1):165–175.

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

Yves Scherrer and Achim Rabus. 2019. Neural morphosyntactic tagging for Rusyn. *Natural Language Engineering*, 25(5):633–650.

Janine Siewert, Yves Scherrer, Martijn Wieling, and Jörg Tiedemann. 2020. LSDC - a comprehensive dataset for Low Saxon dialect classification. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, page 25–35, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Wietse de Vries, Martijn Bartelds, Malvina Nissim, and Martijn Wieling. 2021. Adapting monolingual models: Data can be scarce when language similarity is high.

[10] https://github.com/Helsinki-NLP/LSDC-morph