The Low Saxon LSDC Dataset at Universal Dependencies

Janine Siewert, Jack Rueter
University of Helsinki
Department of Digital Humanities
janine.siewert@helsinki.fi, jack.rueter@helsinki.fi

Abstract

We present an extension of the Low Saxon Universal Dependencies dataset and discuss a few annotation-related challenges. Low Saxon is a West-Germanic low-resource language that lacks a common standard and therefore poses challenges for NLP. The 1,000 sentences in our dataset cover the last 200 years and 8 of the 9 major dialects. They are presented both in original and in normalised spelling and two lemmata are provided: A Modern Low Saxon lemma and a Middle Low Saxon lemma. Several annotation-related issues result from dialectal variation in morphological categories, and we explain differences in the pronoun, gender, case, and mood system. Furthermore, we take up three syntactic constructions that do not occur in Standard Dutch or Standard German: the possessive dative, pro-drop in pronominal adverbs, and complementiser doubling in subordinate interrogative clauses. These constructions are also rare in the other Germanic UD datasets and have not always been annotated consistently.

Keywords: Universal Dependencies, Low Saxon, Low German, low-resource, non-standard

Figure 1: The Low Saxon language area and its major dialect groups.

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

Table 1: The size of the dataset by the dialects included.

1. Introduction

Low Saxon[1] is a West-Germanic language primarily spoken in the north-eastern Netherlands and northern Germany by an estimated number of 3–4 million people. Figure 1 shows the area in the early 20th century including the East Pomeranian (POM) and Low Prussian (NPR), which were spoken in these eastern areas until the end of WWII.

Despite official recognition in both countries, there is no interregional standard. This means that a high degree of both orthographic and dialectal variation needs to be taken into consideration in NLP. Since carefully annotated datasets for Modern Low Saxon are still scarce, this extension of the UD dataset is part of our ongoing efforts to fill this gap.

2. Data

Table 1 shows the size of the dataset by the dialects included. The East Pomeranian (POM) dialect shown in Figure 1 is missing due to a lack of data, since the only resources we have managed to obtain so far consist of poems and songs which do not lend themselves well to syntactic annotation. Our dataset primarily covers the last 200 years with a focus on the older period, since it is easier to obtain copyright-free material from that time. In addition, due to data scarcity, we have included a few sentences from the late 17th century in the Eastphalian (OFL) part of the dataset.

The largest part of the new sentences stems from train, develop and test sets we had annotated for other purposes. These train, development and test sets had been randomly chosen from the LSDC dataset (Siewert et al., 2020), which has led to some imbalance in the dialect representation. When we complemented this data with other resources to raise the number of sentences to 1,000, we prioritised the underrepresented dialects. This means an addition of more than 900 sentences compared with the 2.12 release from May 2023.

The sentence ID provides information on the dialect group and year of publication. The first segment is the corpus name 'LSDC', the second one the number of the sentence, the third segment the abbreviation of the dialect group (e.g., 'NNS'), and the fourth segment the year of publication. These

[1] Also called 'Low German'

LREC-COLING 2024, pages 15976–15981

might be followed by further information on, for instance, the dialect subgroup, the exact place of publication, the name of the author or the title of the work, where available.

The preannotation of PoS tags, morphological features, and dependency relations was done with Stanza (Qi et al., 2020) trained on West Germanic and mainland Scandinavian UD datasets, while the Stanza lemmatiser was trained on the Reference Corpus Middle Low German / Low Rhenish (ReN-Team, 2021). Subsequently, we have manually corrected and validated these automatic annotations. The annotation of the main lemmata in Modern Low Saxon has been an entirely manual undertaking.

3. Spelling normalisation and lemmatisation

The sentences are provided both in the original spelling in the `# text_orig` line, and in normalised spelling in the `# text` line.

In the absence of a common standard, we use the interregional spelling *Nysassiske Skryvwyse*[2] that has been adopted by the Dutch Low Saxon Wikipedia and the Low Saxon Wiktionary. This spelling strives to bring the Low Saxon dialects closer together orthographically by basing the grapheme usage on historical phoneme correspondences instead of particular local modern pronunciations. For instance, the word for *book* that might be written as *Bauk*, *bouk*, *book*, *Bok*, *boek* or *beok* in a variety of local spellings is unified to *book* based on the origin of the vowel in Proto-Germanic *ō. Nevertheless, the *Nysassiske skryvwyse* represents variation to account for divergent developments that make it impossible to regularly derive all local pronunciations from one orthographical form. The cognate of *over*, for example, may be written *oaver* or *öäver* depending on whether the vowel exhibits i-mutation in the dialect in question. The normalised text reflects such local characteristics.

The main lemma is given in the *Nysassiske Skryvwyse* as well, but here, we use the same lemma forms across dialects. In instances where divergent forms are found in today's dialects, the variant closest to the Middle Low Saxon dictionary form is chosen as the modern lemma. The North Sámi UD corpus (Sheyanova and Tyers, 2017) follows a similar approach with representation of dialectal variation in the `# text` field and a standardised lemma.

In some cases, however, more than one modern lemma can correspond to the same Middle Low Saxon lemma. This usually happens when a word has developed different forms dependending on the syntactic function. For instance, the indefinite and definite article have developed forms distinct from the corresponding indefinite and demonstrative pronouns in many dialects. Therefore, we lemmatise the articles as *en* and *de*, and the pronouns as *eyn* and *dee*.

In addition to Modern Low Saxon lemmata, we provide a second lemma in normalised Middle Low Saxon in the tenth column as `lemma_gml`, since linguists working on (historical) Low Saxon are likely to be familiar with this spelling. Our spelling largely follows the *Mittelniederdeutsches Handwörterbuch* by Lasch et al. (1928 ff.) that is also used in the Reference Corpus Middle Low German / Low Rhenish (1200–1650) (ReN-Team, 2021). In order to ease the manual annotation, however, we have removed superscript numbers and parantheses, and made a few grapheme simplifications such as ‹êⁱ› to ‹êi› and ‹a̋› to ‹a› or ‹ā› depending on the context. Furthermore, we have unified the annotation of certain suffixes: E.g., the usage of the variants *līk*, *-līke*, *-līken*, *-līke(n)*, *-līke(s)*, *-liken* and *-haftich*, *-hachtich*, *-haft*, *-hacht*, *-haftigen*, *-aftich* was systematised in such a way that *-līk* and *-haftich* are used for the adjective and *-līken* and *-haftigen* for the adverb.

4. PoS and Morphological annotation

The Low Saxon dataset uses all Universal PoS tags[3] except for 'SYM'. So far, this tag has not been necessary, as the dataset does not contain emojis, currency symbols or mathematical operators yet.

Due to the lack of a common standard, we find a few differences in morphological categories in the various dialects and time periods. This means that the same form can fulfill different functions based on the variety, which is reflected in the annotation of morphological features.

4.1. Personal pronouns

The reference of some pronouns is ambiguous due to (historical) use as expressions of politeness.

Several Dutch Low Saxon dialects, especially in the southern part of the language area, have lost or are in the process of losing the second person singular pronoun *du/dû* and extending the use of *jy/gî* to the singular. It is often not possible to tell from the context whether a single person is addressed with *jy/gî* for reasons of politeness or if the pronoun *du/dû* has already fallen out of use in the variety in question. Nor is it always possible to decide whether one or more persons are being referred to by *jy/gî*.

The annotation of *jy/gî* and verbs agreeing with it follows the context. Thus, the annotation

[2] A more detailed description in Low Saxon can be found here: https://skryvwyse.eu

[3] https://universaldependencies.org/u/pos/index.html

is `Number=Sing` for a single person and `Number=Plur,Sing` if reference is ambiguous. Politeness annotation `Polite=Form` is only added if the context leaves no room for doubt and in our dataset we have only encountered such cases in northern Dutch Low Saxon and German Low Saxon.

In German Low Saxon, we find the third person plural *see/sê* used as a politeness marker in part of the dialects. Since the polite pronoun and agreeing verbs refer to the addressee, they will be annotated as `Person=2`.

4.2. Feminine-masculine distinction

Another difference in morphological categories is the number of grammatical genders. Most Low Saxon dialects have three genders: feminine, masculine, and neuter, whereas remainder see the merging of the feminine and masculine genders into a common gender and show no overt distinction on the noun, agreeing determiners or adjectives. Personal pronouns, however, might reveal the gender of the noun.

Gender marking poses problems for both the human annotator and the annotation model for two other reasons as well: First, even in dialects with distinct feminine and masculine forms, the distinction is not overt in all inflectional forms. In many dialects, for example, feminine and masculine noun phrases have the same form in the determinate nominative, while forms are distinct in the indeterminate nominative or the determinate accusative. Additionally, no gender distinction is marked in the plural. Second, there is dialectal variation in gender assignment and, for many local varieties, the gender of words is not documented. This is especially true for Dutch Low Saxon, where many dictionaries do not even mention the gender category. This situation is further complicated by the contact with Dutch. As Bloemhoff et al. (2019) have observed, speakers of Twents, a Dutch Westphalian dialect, increasingly struggle with the gender assignment of nouns that have common gender in Dutch.

Generally, we strive to disambiguate the gender based on the context or other available sources, such as dictionaries for the target or a closely related variety. If no such information can be obtained, or the noun can carry both genders, we use the combined feature `Gender=Fem,Masc`. This feature is also used for gender-ambiguous forms of pronouns, such as the demonstrative/relative pronoun *dee/dê* 'that' or the interrogative pronoun *wek/welk* 'which'.

4.3. Case inventory

The size of the case system differs from dialect to dialect. While nominative and accusative are distinguished in the personal pronouns everywhere (compare Lindow et al., 1998), especially varieties that have been influenced by Dutch do not inflect nouns for case. In addition to Dutch Low Saxon, this is true for, e.g., East Frisian[4]. (Lücht, 2016, 62)

Despite the lack of overt marking in a few varieties, the nominative-accusative distinction is annotated throughout on pronouns, nouns, and agreeing dependents such as determiners and adjectives.

Productive dative forms have been preserved in several German Westphalian and Eastphalian varieties (cf. Lindow et al., 1998), as shown in Example (1) from East Westphalian.

(1) Ik sto in der Gemoene iarem
I stand in the.DAT parish.DAT her.DAT
Denste
service.DAT
'I stand in the service of the parish.'

In other areas, dative remnants may occur after specific prepositions, e.g., in the German North Saxon Example (2) after *bi*. If the regular accusative form were used, the result would be *bi 't Läsen*.

(2) Mi weer de Sunn to grall bi
me was the sun too bright at
'n Läsen .
the.DAT.SG reading .
'The sun was too bright for me while reading.'

The historical genitive, however, has almost completely fallen out of use in the whole language area and Example (3) exhibiting this case, in fact comes from the late 17th century Eastphalian texts mentioned in Section 2.

(3) na synes öldesten Broders
after his.GEN oldest.GEN brother.GEN
Doode
death.DAT
'after the death of his oldest brother'

Forms bearing the feature `Case=Gen` most commonly refer to a possessive construction resembling the English "Saxon genitive", such as in the Dutch North Saxon example *'n mouders ooge zucht scharp* 'a mother's eye sees sharp' from our dataset.

4.4. Subjunctive mood

Distinct subjunctive forms have been lost in most Low Saxon dialects today, and their use is mostly confined to parts of Westphalia and Eastphalia. One predominantly finds the past subjunctive in

[4]The varieties spoken in East Frisia which are part of German North Saxon (DNS); not the East Frisian language Saterlandic spoken outside of East Frisia.

these areas, as in the Westphalian Example (4) from Saltveit (1983, 299), while the present subjunctive is rare everywhere and likely restricted to a few fixed expressions.

(4) et söl mi frögn, wank et
it shall.PST.SBJV.3SG me please if-I it
bekäme
get.PST.SBJV-1SG
'I would be happy if I got it.'

In dialects without distinct past subjunctive forms, the past indicative can be used instead, as can be seen in the German North Saxon Example (5) from Saltveit (1983, 300).

(5) du schusst man lewer to Huus
you.SG shall.PST-2SG but rather to house
gahn hebben
go have
'You had better gone home.'

An interesting case is found in many northern Low Saxon dialects in Germany, where past subjunctive forms have replaced the past indicative in certain inflectional classes. As a result, subjunctive forms are used without distinction in subjunctive and indicative meaning. (Saltveit, 1983, 298–301).

Formally ambiguous cases, i.e., indicative forms in contexts that historically or in other dialects would require subjunctive forms are annotated with the combined feature `Mood=Ind,Sub`.

5. Dependencies

5.1. Possessive dative construction

Low Saxon has three primary possessive constructions: A "Saxon genitive" similar to English, a prepositional construction using *van* 'of', and a possessive dative. In the possessive dative construction, the possessum is preceeded by the possessor in the dative case[5] and a possessive determiner, as can be seen in Example (1) above.

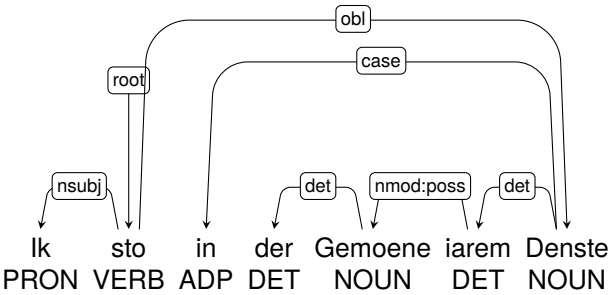

Figure 2: Possessive dative

[5] In case a distinct dative case has been preserved, otherwise a form corresponding to the accusative or nominative may appear.

Among the UD languages, we have found comparable constructions in Afrikaans, Frisian Dutch, and Norwegian, but the annotation has been inconsistent across these languages.

As shown in Figure 2, the possessive determiner connects to the head noun with a *det* relationship, and the possessor with an *nmod:poss* relationship to the possessive determiner of the possessum. The reason for this is that the possessive determiner of the possessum is obligatory, so that the possessor in the dative case cannot be connected without it.

5.2. Pro-drop in separable pronominal adverbs

Similar to Dutch, pronominal adverbs in Low Saxon are commonly separable. We follow the Dutch annotation and connect the pronominal part tagged as 'ADV' as `obl`, and the prepositional part tagged as 'ADP' is connected to it as `case`.

This is illustrated by the German Westphalian sentence in Figure 3 from our dataset.

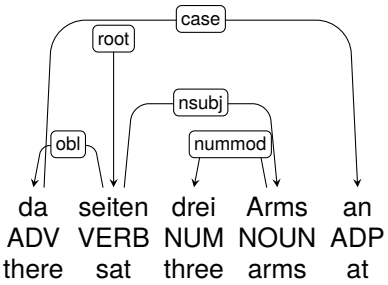

Figure 3: Separable pronominal adverbs; 'three arms were attached to it'

Unlike in Standard Dutch, pro-drop of the pronominal part is fairly common, leaving only the prepositional part. An example of this is the Brandenburgish sentence in Figure 4 from our dataset.

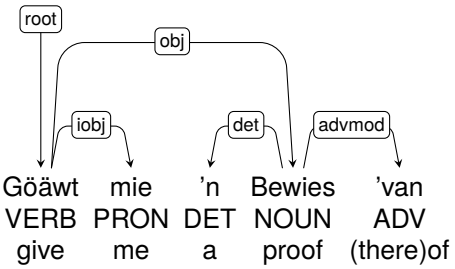

Figure 4: Pro-drop in pronominal adverbs

In such cases, we tag the remaining prepositional part as 'ADV' and connect it as `advmod`, in the same way an unseparated *doavan* would be.

5.3. Complementiser doubling in subordinate interrogative clauses

In several Low Saxon varieties, interrogative adverbs in indirect questions as well as relative pronouns can be followed by an additional complementiser *as* 'as' or *dat* 'that'. This is illustrated by the German North Saxon sentence from Wisser (1921, 29): *Un dạrmit secht de ol Mann em Beschêd, wodenni as he dat maken schall.* 'And with this, the old man tells him how he should do it.' This construction is already attested in Medieval Low Saxon and occurs in other West-Germanic varieties as well (Schallert et al., 2018), for instance, in Frisian (Popkema, 2018, 299). In contrast, this construction does not occur in either of the majority languages Standard Dutch or Standard German.

This construction appears in the Frisian Dutch UD dataset (Braggaar and van der Goot, 2021), where they unexpectedly write it as a contraction instead of splitting it on two lines. E.g, in *de order dy't no binnenkaam is* 'the order that has now come in', *dy't* is treated as a single element.

In contrast, the following sentence from Wisser (1921) shows that, at least in Low Saxon, other words can be placed in between the interrogative adverb (*wo* 'how') and the complementiser *as*: *un will [...] sêhn, wo wid as se tô sünd* 'and wants to see how far they are'. As a result, we connect it as `mark` to the verb and not as `fixed` to the adverb as shown in Figure 5.

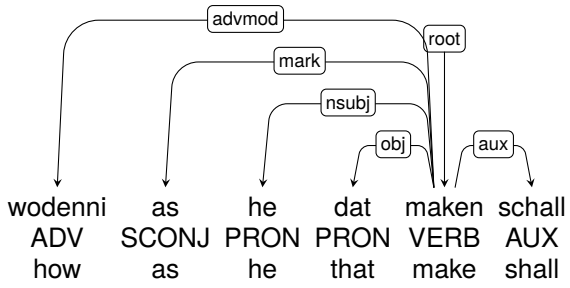

Figure 5: Complementiser doubling.

6. Summary and future plans

We have presented a substantial extension of the first Low Saxon dataset on Universal Dependencies and hope our discussion of approaches for dealing with the lack of a common standard variety to be useful for annotation work on other languages in comparable situations.

We plan to continue the extension of the treebank and work in particular on the underrepresented varieties. This will likely require us to also draw from other sources than the original LSDC by Siewert et al. (2020).

Furthermore, we would like to initiate an exchange on the annotation of, e.g., the possessive dative construction in order to come to an agreement across datasets and languages.

7. Acknowledgements

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

8. Bibliographical References

Henk Bloemhoff, Philomèle Bloemhoff de Bruijn, Jan Nijen Twilhaar, Henk Nijkeuter, and Harrie Scholtmeijer. 2019. *Nedersaksisch in een notendop – Inleiding in de Nedersaksische taal en literatuur*. Koninklijke Van Gorcum, Assen.

Anouck Braggaar and Rob van der Goot. 2021. Challenges in annotating and parsing spoken, code-switched, Frisian-Dutch data. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 50–58, Kyiv, Ukraine. Association for Computational Linguistics.

Agathe Lasch, Conrad Borchling, Gerhard Cordes, Dieter Möhn, Ingrid Schröder, Jürgen Meier, and Sabina Tsapaeva. 1928 ff. *Mittelniederdeutsches Handwörterbuch*. Wachholtz Verlag, Neumünster.

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

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

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

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and 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*.

Laurits Saltveit. 1983. Syntax. In Gerhard Cordes and Dieter Möhn, editors, *Handbuch zur niederdeutschen Sprach- und Literaturwissenschaft*, pages 279–333. Erich Schmidt, Berlin.

Oliver Schallert, Alexander Dröge, and Jeffrey Pheiff. 2018. Doubly-filled COMPs in Dutch and German – A Bottom-up Approach. *Universities Munich and Marburg, Manuscript*.

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

9. Language Resource References

ReN-Team. 2021. Reference Corpus Middle Low German/Low Rhenish (1200–1650); Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200–1650).

Mariya Sheyanova and Francis M. Tyers. 2017. Annotation schemes in north sámi dependency parsing. In *Proceedings of the 3rd International Workshop for Computational Linguistics of Uralic Languages*, pages 66–75.

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*, pages 25–35, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).