Towards the Morphological Annotation of North Markian (Low German)

Christian Chiarcos
Applied Computational Linguistics (ACoLi)
University of Augsburg, Germany
christian.chiarcos@uni-a.de

Abstract

Low German (Low Saxon, ISO 639-2 `nds`) is an underresourced West Germanic language spoken in Northern Germany (*Plattdütsch*), in the Netherlands (*Nedersaksisch*) and in an international diaspora (*Plautdietsch*, *Pomerano*, etc.). As a minority language, it is under pressure from the respective national languages, and considered threatened. Although NLP and digital language resources might play a role in facilitating the use of the language on the web and to support intergenerational transmission, no NLP tools are known to exist, and no adequate corpora that such tools could be trained on. This paper describes a novel corpus of North Markian, a dialect of East Low German, its morphosyntactic annotation and morphological analysis, and in particular explores methods to bootstrap and develop such resources in the face of a complete lack of training data.

Keywords: Low German (Low Saxon), underresourced languages, morphology

1. Background and Motivation

Low German (Low Saxon) is a West Germanic language primarily spoken in Northern Germany and the Netherlands, most closely related to German, Dutch and Frisian, and, as the language of the Hanseatic League, highly influential in Scandinavia and the Baltics during the Middle Ages. Despite its name, Low German is not a dialect of High German, but a sister language with an independent literary tradition of more than a millenium. It is a group of West Germanic dialects descending from Old Low German (Old Saxon) and later Middle Low German, spoken in Northern Germany and neighboring regions since the Migration Period (ca. 500 CE) and throughout the Middle Ages. Modern Low German dialects share several innovations, such as the reduction of nominal and verbal morphology, while lacking characteristic developments of neighboring Germanic branches: the High German consonant shift, Anglo-Frisian palatalization, and Dutch diphthongization (see Table 6). The decline of Low German as a written language followed the military defeats of the Hanseatic League in the 16th century, the devastation of the Thirty Years' War (1618–1648), and the later integration of its speech area into the emerging national states of Prussia/Germany and the Netherlands, which promoted High German and Dutch as administrative and educational languages. Today Low German is characterized by widespread multilingualism and strong influence from the respective national languages.

In terms of resources available for natural language processing and corpus linguistics, modern Low German is highly underresourced. Despite an extensive (but mostly regional) literature and considerable dialectological research in these varieties during the 19th and 20th c., this includes a general lack of machine-readable dictionaries (Chiarcos et al., 2025a) and a deficit of linguistically annotated corpora (see Section 4).

In this paper, we describe the development of a corpus of North Markian (further NMk) and its morphosyntactic and morphological analysis. Within Eastern Low German, the North Markian dialects represent a continuum of varieties (Figure 1) that are spoken in three federal states of Germany, Saxony-Anhalt (Altmark), Brandenburg (Prignitz, Uckermark), and Mecklenburg–Western Pomerania (Western Pomerania). The latter variety is also referred to as 'Central Pomeranian' (Pfaff, 1898), but in its characteristics largely identical with North Markian and thus included in our corpus. North Markian and Central Pomeranian are defined by a number of common traits that set them apart from Mecklenburgian/Western Pomeranian (further, MWP) to the north and Central Markian (CMk) to the south. The defining criterion of NMk within the Eastern Low German dialect group is the development of Middle Low German long vowels (Table 1). Other common features are a distinctive stock of Markian vocabulary (shared with CMk, largely inherited from Low Franconian/Dutch settlers from the Middle Ages (Teuchert, 1944)), and the apocopy of unstressed Middle Low German syllables (shared with MWP, but different from CMk). This has had far-reaching effects in the areas of morphology and probably also syntax, although both aspects are poorly covered by scientific literature and remain largely unexplored due to the lack of commonly available empirical data.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11699–11714
11–16 May 2026.

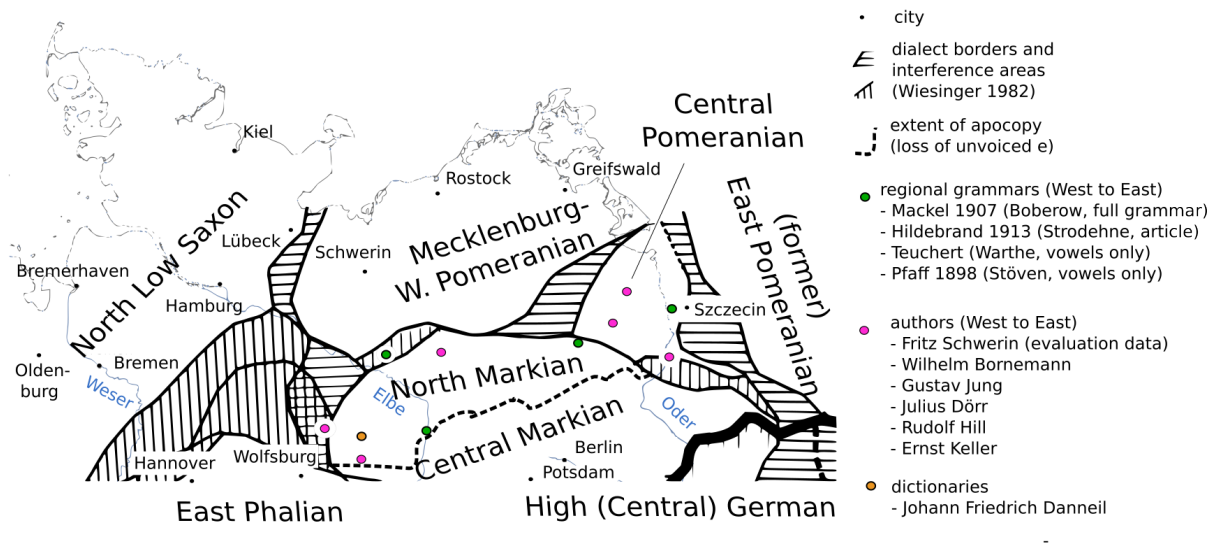

Figure 1: North Markian, Central Pomeranian and neighboring Low German (North Low Saxon, East Phalian, Central Markian, Eastern Pomeranian, Mecklenburg-Western Pomeranian) and High German dialects (Central German) in Northern Germany

MLG	MWP	NMk	CMk	German	Dutch	English
ê¹	/e:/ *Kes'*	/e:/ /ke:z/	/ɛ:/ /kɛ:zə/	*Käse*	*kaas*	"cheese"
ê²	/aᴵ/ *Bein*	/e:/ /be:n/	/ɛ:/ /bɛ:n/	*Bein*	*been*	"leg"
ê⁴	/aᴵ/ *Breiw*	/e:/ /bre:f/	/iᵉ/ /briᵉf/	*Brief*	*brief*	"letter"
ê³	/aᴵ/ *Arbeit*	/aᴵ/ (/ɛi/) /arbaᴵt/	/aᴵ/ /arbaᴵt/	*Arbeit*	*arbeid*	"work"
ô¹	/aᵘ/ *dauhn*	/o:/ /do:n/	/uᵒ/ /duᵒn/	*tun*	*doen*	"to do"
ô²	/o:/ *Brod*	/o:/ /bro:t/	/ɔ:/ /brɔ:t/	*Brot*	*brood*	"bread"

Table 1: Sound correspondences between Middle Low German (Seelman, 1908, MLG), Mecklenburg-West Pomeranian (Müller, 1904, MWP, as written by Fritz Reuter), North Markian (NMk) and Central Markian (Seelman, 1908, CMk).

This situation poses particular challenges for the preservation and teaching of NMk, because it means that didactic materials, dictionaries and texts from other Low German varieties can only be used to a limited extent to convey language skills. This situation is particularly problematic as long as the dialect-specific features (beyond phonology, which has been well researched by a number of local grammars from the 19th and 20th centuries) remain undocumented, making it difficult for teachers and learners to correctly contextualize materials and texts from other Low German varieties and to assess aspects in which the local vernacular is expected to deviate in a systematic manner. There is currently no dictionary of North Markian in existence,[1] The situation is further complicated by the author-specific orthographies of the available texts, which are usually based on High German, so that they do not adequately reflect the Low German differentiation between short, lengthened and long vowels and are therefore systematically ambiguous.

To counteract this deficiency, we have started preliminary work on building a morphosyntactically annotated corpus of North Markian. Our design goals are pragmatic in that we aim to minimize manual annotation costs, ensure traceability from normalized forms back to the page image, and enable portability to neighboring Low German dialects. We adopt finite-state methods because they offer transparent, linguistically interpretable analyses and well-understood engineering trade-offs (Beesley and Karttunen, 2003). In contrast to end-to-end neural methods, finite-state morphology remains competitive for low-resource settings where expert knowledge and limited data can be combined effectively.

The modern Low German language is highly fragmented, both politically, but also linguistically and orthographically. The most realistic approach is thus to not propose one solution for the language as a whole, but to focus on regional varieties, and

[1]Commonly available dictionaries (Danneil, 1859; Stellmacher, 1996; Kettmann, 2004) are defined by *regions*, not dialects, and all these dictionary lump together both North Markian and other varieties in the respective regions (High German dialects, CMk, East Phalian).

then to use automated transliteration routines to facilitate the transfer between different orthographies and dialects. So, we focus on one particular variety, only, and use North Markian for illustration, as this is a particularly poorly resourced variety of Low German for which no designated dictionaries and only one more extensive grammars (Mackel, 1905, addressing a transitional dialect between MWP and NMk) are known to exist, but which nevertheless extends across a relatively wide area within Germany. We develop (1) a workflow for digitizing, normalizing, and annotating NMk texts; (2) a computational morphology for inflection and word formation; (3) a lemma list and a full form dictionary of North Markian, and (4) workflows and templates that can be re-used for other low-resource varieties. Beyond North Markian, this allows to assess the potential of the technologies we used for other varieties of Low German, and, in the longer perspective, develop similar solutions (and, hopefully, some level of reusability) for the language as a whole, in Germany, the Netherlands and beyond.

2. Digitization and Initial Morphosyntactic Annotation

So far, digitized text in North Markian is only available in small excerpts,[2] whereas more extensive material is digitally available in the form of scanned books (i.e., images, not text). From this pool of data, we initially identified copyright-free texts that reflect the lexics and morphology of five main regions, with Bornemann (1810, 1816, 1868) for Altmark, Jung (1855) for Prignitz, Dörr (1888) for Uckermark, Hill (1868) for Central Pomeranian (Northern Uckermark), and Keller (1871, 1872, 1877) for varieties formerly spoken in modern Poland (Table 2). The aim of this initial selection was to compile a common word list in phonemic spelling in order to develop tools for the morphological and morphosyntactic analysis of North Markian and its varieties, and thus to provide the basis for NLP and the corpus-linguistic coverage of North Markian. Unfortunately, for copyright reasons, only a very limited number of prose texts are available (of the authors mentioned, only Dörr is not poetic in nature), which is why the analysis we provide so far has an inherent focus on lexis and morphology, whereas future consideration of syntax requires the inclusion of texts that either have yet to be digitized or for which copyright clearance is still pending.

All texts are set in Gothic script (Fraktur), and are accessible online (e.g., via Google Books) with

author	genre	variety	region	tokens
Bornemann	poems	NMk	Altmark	87.747
Jung	poems	NMk	Prignitz	33.618
Dörr	novel	NMk	Uckermark	98.268
Hill	poems	CP	Uckermark	23.339
Keller	poems, play	CP	Poland	37.673*
total				280.645

Table 2: Composition of the North Markian corpus. Note that Central Pomeranian (CP) is considered here a variety of North Markian, but formally distinguished from North Markian proper (NMk)

poor-quality OCR, only. Our texts are 298the result of automatic text recognition with Transkribus (Kahle et al., 2017, model Print M1, version of February 19, 2022), which supports modern and historical German, Dutch and Swedish, among others, and thus covers the range of author-specific diacritics relatively well. We also performed comparative evaluation with German Kraken models,[3] which did, however, perform generally slightly worse. Note that we did not perform any OCR correction. Instead, OCR errors are dealt with in a subsequent normalization step.

For the 1000 most common word forms of each source text, we manually identified all possible morphosyntactic annotations (parts of speech according to Universal Dependencies De Marneffe et al., 2021, UD). Furthermore, a list of candidate verbs, adjectives and nouns was extracted from the corpus using heuristic filters. The resulting list of word forms and possible parts of speech was further enriched with lemmas from dictionaries of related varieties (Dictionary v0.1, see below) and used to estimate the emission probabilities of a Hidden Markov Model for POS annotation, which were combined with the transition probabilities of several UD-annotated corpora (maximum likelihood estimation, no smoothing; Dutch: Alpino; German: HDT, GSD, LIT, PUD; Swedish: PUD, Talbanken) and corpora of historical varieties heuristically mapped to UD parts of speech (Barteld et al., 2017 for Middle Low German; Dipper, 2015 for Middle High German) to automatically generate morphosyntactic annotations. Figure 2 summarizes these steps.

The resulting annotations were evaluated against a small, manually annotated test set consisting of the text from the respective page 100 of Bornemann 1816 and 1810, Dörr, Jung and Hill, a total of 950 tokens. The best-performing configuration yielded an accuracy of 91.9% (Dutch, Alpino, case insensitive, no retraining) on the test corpus, with precision and recall per POS as shown in Table 3. Overall, the model struggles specifically to disambiguate adjectives and participles, as well as nouns and proper

[2]The largest data set we are aware of are 11 poems by Wilhelm Bornemann, https://wikisource.org/wiki/Bornemann:_Plattdeutsche_Gedichte.

[3]https://kraken.re/main/index.html

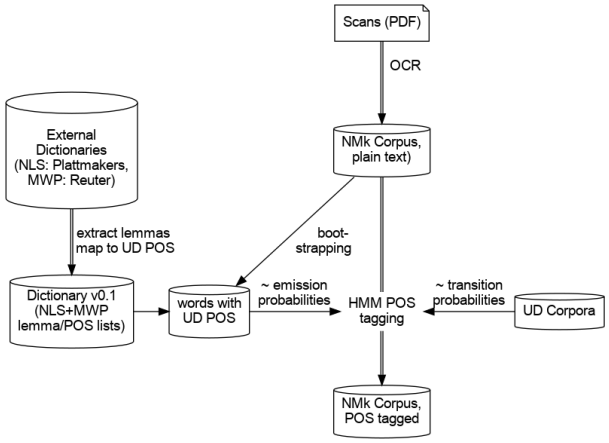

Figure 2: Workflow to create a POS-tagged corpus without training data

tag	prec	rec	f
CCONJ	96.3%	100%	98.1%
DET	97.2%	98.6%	97.9%
AUX	100%	95.5%	97.7%
PRON	96.0%	96.0%	96.0%
ADV	98.1%	82.8%	89.8%
ADP	84.2%	96.0%	89.7%
NOUN	87.3%	92.1%	89.7%
VERB	79.5%	96.8%	87.3%
SCONJ	100%	69.2%	81.8%
PROPN	91.7%	68.8%	78.6%
ADJ	61.1%	64.7%	62.9%

Table 3: Precision, recall and f-score of the the best-performing HMM tagger (Dutch, Alpino, lower-cased, no retraining (transition probabilities directly from Alpino corpus).

nouns. Overall, while differences between the models were marginal, the good performance of the Dutch model is not unexpected. Although a sociolinguistic perspective (Stellmacher, 2017) might have predicted a High German model to perform better on our data, Low German and Dutch share a number of characteristics in their grammar that set them clearly apart from German. Aside from aspects of diachronic phonology, this specifically includes apocopy and the wide loss of case morphology that comes along with it. Moreover, Markian dialects were historically strongly influenced by Low Franconian, resp. Middle Dutch varieties – to the extent that Markian was occasionally described in earlier literature as the language of medieval Dutch colonies (Teuchert, 1944). Given the small size of the test set, however, an element of chance seems likely.

3. Developing a Finite State Morphology and Bootstrapping its Lexical Backbone

For POS tagging, we could exploit a key property of HMMs: the separation of emission probabilities (linking word forms to POS tags) from transition probabilities (modeling POS sequences independently of word forms). This allowed us to combine foreign-language transition probabilities with language-specific emission probabilities heuristically derived from bootstrapped word lists.

For higher levels of analysis, however, such transfer is not feasible. Low German in general, and North Markian in particular, lack not only direct training resources but also the parallel corpora or dictionaries needed for transfer learning. To the best of our knowledge, attempts to use LLMs for Low German have been reported as unsuccessful (see see the limitations section of this paper), and also our own experiments on transliteration (see limitations) and zero-shot morphosyntactic annotation (not reported here) led to disappointing results, so far. Adapter-based approaches are equally difficult, because North Markian morphology and syntax remain insufficiently described: before fine-tuning a model, the target analysis must first be established. This makes corpus annotation an exploratory task as much as a technical one. We operate with the following working hypotheses: North Markian is assumed to be relatively uniform in phonology; its synthetic morphology is treated as preserving at least part of the Middle Low German category system, including accusative vs. dative and indicative vs. subjunctive; its lexicon is expected to overlap with neighboring Low German varieties, especially Mecklenburgian and North Low Saxon; and Mackel (1905), despite describing a transitional Mecklenburgian dialect, provides a useful starting point for morphology.

Starting from these assumptions, we adopt an iterative workflow of computational analysis, manual evaluation, and hypothesis revision. From an initial phonological model and external dialect resources, we bootstrap North Markian morphology and lexical resources from the corpus itself, primarily using symbolic methods, especially lemma lists and FSTs. The overall workflow is summarized in Figure 4 in the Appendix.

Borrowing external dictionaries (aka "Dictionary v0.1") It is also to be noted that there is no digital (or digitized) dictionary of any North Markian variety that could serve as a basis for developing a computational morphology. There are, however, dictionaries of other Low German dialects for which a considerable overlap in core vocabulary can be expected. Here, we specifically looked

into the dictionaries of Buck (2007-2024, Plattmakers), a dictionary with a focus on North Low Saxon (further NLS), and Müller (1904), a dictionary of the writings of Fritz Reuter (Mecklenburgian, further MWP). The first was chosen because it provides a phonological transliteration, from which the actual phonemes in NMk could be partially predicted, whereas most other dictionaries use more or less deficient orthographies. The second dictionary was chosen because of the influence that Fritz Reuter executed on subsequent literature in Eastern Low German, and because its language and the material culture it describes is contemporary to those of our sources. Both are available over the internet, but not in machine-readable form, so that we devised a crawler to extract word forms and their parts of speech from the digital sources. We provide both dictionaries and their normalization to North Markian as part of preceding research (Chiarcos et al., 2025b). In addition to their phonological normalization as described below, these have been used to extract lists of lemmas and parts of speech: Both dictionaries comprise partial part-of-speech information, albeit only according to the specifications of the authors. Where available, these were mapped to UD by the authors of the current paper.

Phonology and Normalization (aka "Lemmatizer v0.0") Using the initial POS annotation for filtering, we identified base forms of verbs, nouns and adjectives from our corpus. A subset of these were matched with the NLS and MWP lemma lists, with the goal to develop an initial phonological normalization. It is to be noted that there is no standard orthography for Low German. The various orthographies used by individual authors and in the dictionaries we operate with thus deviate substantially from each other (see in Table 7 in the appendix). Differences include, for example, the distinction between long closed, long open and short vowels (e.g., in the use of unambiguous *ao* or *oa* alongside ambiguous *a*, *o* for /ɔː/), writing conventions for assimilations (e.g., of historical *-ben* in historical *hebben* 'to have' or of intervocalic *-dd-* in historical *hadden* '(we) had'), clitics (sometimes marked by apostrophe, but not in *harnw* 'had we'), apocope and syncope (sometimes marked by apostrophe, e.g., in *Fru'ns* alongside *Fruns(lüd)* 'women'). All existing orthographies are highly individual, as the authors employed different strategies to compensate the deficits of German-based orthographies in application to Low German. But this also means that spelling variation across different authors can help to disambiguate orthographic ambiguities. In order to preserve as much information about the underlying phonology as can be retrieved from the authors' orthographies, we introduce both a phonological normalization that allows to abstract over the different orthographies, and author-specific routines for the mapping of their orthographies to this normalization. Internally, we represent NMk phonology in a normalized way with the following symbols: *a,e,i,o,u,ö,ü* for short vowels, *E,I,O,U,Ö,Ü* for long closed vowels, *au,ei,äu* for diphthongs, and *Ä,œ,å* for the long open vowels /ɛː/, /œː/ and /ɔː/; the consonants (*b,d,f,g,h,j,k,l,m,n,p,r,s,t,v,w,x,S*) include *x* for /ç/ and /χ/ and *S* for /ʃ/. Using the Stuttgart FST library (Schmid, 2005, SFST), we then developed two FSTs for normalizing both dictionaries against our phonological representation and 6 FSTS for the individual orthographies of each individual author to this simplified phonological representation. For every dictionary d and every author a, we can then build an FST for word lookup by concatenating an FST for author normalization ($\text{FST}_{a\mapsto norm}$) with (the inverse of) an FST doing dictionary normalization ($\text{FST}^{-1}_{d\mapsto norm}$) and (the inverse of an FST compiled from) the words of the dictionary (FST^{-1}_{d}):

$$\text{FST}_{a\mapsto d} = \text{FST}_{a\mapsto norm} \circ \text{FST}^{-1}_{d\mapsto norm} \circ \text{FST}_{d}$$

Inflectional Morphology based on Mackel (1905) (Lemmatizer v0.1) We developed an initial morphology operating on this phonological representation on the basis of Mackel (1905), who provides a description of the inflectional morphology of the dialects of the Prignitz region. To the best of our knowledge, this is the only scientific description that – partially, at least – addressed the morphology of a North Markian variety, but even here, it is not clear whether this is indeed the case: Mackel claimed that (most of) his observations apply to the Prignitz as a whole (so, including local varieties of North Markian), but his reference dialect was a transitional Mecklenburgian dialect. In combination with the normalization FST, this nevertheless allowed to perform full-fledged lemmatization (and morphosyntactic analysis) of North Markian word forms against NLS and MWP lemmas (with the disjunction between the NLS and the MWP transducer, both lemmas are returned, if available):

$$\begin{aligned}\text{FST}_{0.1} \quad = \quad & \text{FST}_{a\mapsto norm} \circ \text{FST}^{-1}_{infl\_m} \\ & \circ \begin{pmatrix} \text{FST}^{-1}_{NLS\mapsto norm} \circ \text{FST}_{NLS} \cup \\ \text{FST}^{-1}_{MWP\mapsto norm} \circ \text{FST}_{MWP} \end{pmatrix}\end{aligned}$$

Note that not just lemmas are returned by $\text{FST}^{-1}_{infl\_m}$, but full-fledged morphological analyses. This is illustrated in Figure 3 for the word *funn* '(he) found', with the normalization `fun`, i.e., /fʊn/, for which the lemma `finen` /fɪnn̩/ 'to find' and (among others) the analysis

VERB.3.Sg.Ind.Prt.st.iii (verb, 3rd person, past indicative, strong inflection, ablaut class III) is proposed. Note that the figure is an illustration of a command-line interface used for disambiguating analyses *produced by* Lemmatizer v0.2 in conjunction with Dictionary v0.2, so that here, already a North Markian lemma is provided and an ablaut class.

Initial NMk lemma list (Dictionary v0.2) Until this point, only very limited resources had been invested in terms of manual, linguistic analysis. More labour-intense has been the subsequent pruning and enrichment process, in which another generation of the FST morphology and the North Markian lemma list have been developed. We used $\text{FST}_{0.1}$ to generated a full form word list for every analyzeable word form attested 10 times or more in our corpus, clustered according to (NLS or MWP) lemmas. These were then manually pruned and NMk base forms were extrapolated from written evidence from our corpus and phonological correspondences with the NLS, resp. MWP lemmas. As an intermediate result of our work, we developed a full-form dictionary covering 2,058 lemmas and 3,657 word forms with 4,349 spelling variants, covering all word forms attested 10 times or more in the corpus. In addition to that, we manually normalized an OCRed version of a dictionary of Low German (NMk and East Phalian) varieties of the Altmark region (Danneil, 1859) to NMk base forms, yielding another 6,158 lemmas.

As a result, we arrived at a word list of North Markian lemmas in normalized phonological representation, albeit with a selection bias, i.e., only lemmas attested for the Altmark (Danneil), Reuter (MWP) or North Low Saxon. However, as part of the reanalysis process, a number of incorrectly lemmatized forms were re-analyzed and the corresponding NMk lemma has been added.

Revised Morphology (Lemmatizer v0.2) Based on observation from the curation of Lemmatizer v0.1 output, the morphology has been revised:

- The implementation of systematic vowel alternations in verbal inflection (ablaut) *as observed in the corpus*.
- Extension of inflection rules with forms and patterns not covered by Mackel (1905).
- A word formation module for derivation and compounding (FST_{wf_m}). This included identifying affixes, their grammatical functions and linking elements used in compounding *based on the corpus*.

The novel lemma list ($\text{FST}_{d0.2}$) then replaced the initial external dictionaries in the lemmatizer:

$$\text{FST}_{0.2} = \text{FST}_{a \mapsto norm} \circ \text{FST}^{-1}_{infl\_m} \circ \text{FST}^{-1}_{wf\_m} \circ \text{FST}_{d0.2}$$

Revised NMk lemma and full-form list (Dictionary v0.3) The revised lemmatizer was applied to the full corpus, resulting in an automatically generated full-form dictionary that was then manually pruned. Every word form for which no analysis had been confirmed during the creation of Dictionary v0.2 was now revisited, with all possible analyses from Lemmatizer v0.3. If only one analysis existed, it was just accepted. If no analysis was proposed, it was to be manually analyzed.

For this pruning process, a simple command-line interface has been designed that lists possible morphological analyses along with all attestations in the respective sub-corpus, shown in Figure 3. Note that it is essential to check attestations along with a lexical entry, as all works considered here also include passages in High German, Yiddish, French or in intermediate vernaculars ('Missingsch'). With this tool, all previously unanalyzed word forms of a source text are processed in lexicographic order such that first, *all* attestations from the source are shown, with the current word highlighted. For the example of *funn* in Jung (1855), only one attestation could be found, *As he nu oaber goar nischt funn* 'but when he didn't find anything, now'. After the attestations, the orthographic form (*funn*) and its phonological normalization (fun for /fʊn/) are shown, then the possible lemmas (here, finen for /fɪnn̩/) and all possible morphological analyses. A morphological analysis consists of a UD part of speech, followed by a .-separated sequence of morphological features. For derived or compound words, the corresponding affix boundaries are marked by - (if a bound morpheme is attached to a word) or = (affix boundary between two bound morphemes), both on the lemma and in the morphological analysis. A user can then check all possible analyses and either add a one manually (key m), change the current analysis (key c), confirm (key +) or drop (key -) the current analysis and continue with the next or skip all remaining analyses (key <DEL> and continue with the next word form. On average, the command line interface allows to prune up to 100 word forms (orthographical variants) per hour.

As a result, we arrived at a near-complete list of all word forms from the corpus along with their analyses and associated lemmas, as well as the associated lemma list. Table 4 summarizes the core statistics of dictionary v0.3.

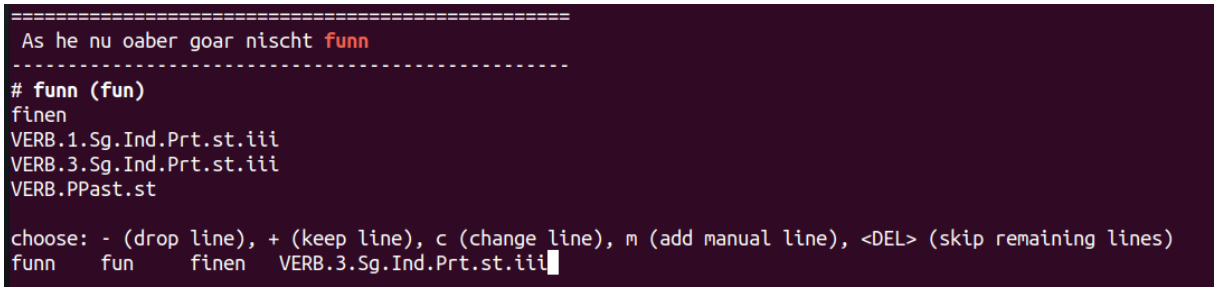

Figure 3: Lemmatizer v0.2+Dictionary v0.2 output for *funn* '(he) found', as shown in the command-line interface for morphological disambiguation

	total	VERB	NOUN	ADJ	ADV	other
forms	17115	7135	5940	2775	2830	2376
normalizations	14361	5985	5901	2395	2196	1582
lemmas	10911	3820	5454	1864	2078	1010

Table 4: Distinct word forms, normalized (phonological) representations and lemmas for the resulting North Markian dictionary (dictionary v0.3)

Cascading Lemmatizer (Lemmatizer v0.3) So far, all lemmatizer components have been specifically set up to work with every individual author and his orthography. For processing unseen text from other authors, it is necessary to increase the band-width of orthographical variation. As an approximation, we set up a generic normalization FST (FST_{norm}) from the disjunction of all author normalization components:

$$\text{FST}_{norm} = \bigcup_{a} \text{FST}_{a \mapsto norm}$$

When using this in a lemmatizer, we found, however, that the additional freedom of variation in the normalization led to an explosion of possible analyses, in particular in conjunction with FST_{wf_m} (because this is recursive), so that retrieving *all* possible analysis was no longer tractable, in many cases. Also, as we had manually confirmed a large number of analyses, a full morphological analysis is no longer necessary if a word form might be already known. We thus implemented a cascading lemmatizer that uses the least labour-intense method for analysis. At present, this is a Python script that calls different FST configurations (if necessary):

given a word form w

return all analysis for w found in the full-form dictionary v0.3 ('cascade 1')

if there are none collect the analyses for every normalization $n \in \text{FST}_{norm}(w)$ that are found in the full-form dictionary v0.3 and return them ('cascade 2')

if there are none return all results of $\text{FST}_{norm} \circ \text{FST}^{-}_{infl\_m} 1 \circ \text{FST}_{d0.3}(w)$ ('cascade 3')

if there are none return all results of $\text{FST}_{norm} \circ \text{FST}^{-}_{infl\_m} 1 \circ \text{FST}^{-}_{wf\_m} 1 \circ \text{FST}_{d0.3}(w)$ ('cascade 4')

In order to achieve real-time performance, the lemmatizer also implements a 1-second timeout as well as a cache to speed up the analysis of recurring words.

Given the lack of independent, manually disambiguated or annotated North Markian (or, generally, Low German) data, we evaluated the resulting analyzer for coverage against a short text by Fritz Schwerin (Schwerin, 1859). Schwerin is an author from the Altmark region, but from its western periphery, writing in a variety with East Phalian influences. We explored two configurations of Lemmatizer v0.3, either with FST_{wf_m} enabled ('cascade 1-3') or disabled ('cascade 1-4') in Table 5. Although both the specific variety (North Markian with East Phalian influences) and the author orthography are unseen, we nevertheless achieve a relatively good coverage of more than 85% of input tokens and about 3/4 of distinct word forms. Manual inspection of a random sample of analyses that involved FST_{wf_m} revealed that in many cases, word formation overgenerated massively, so that the amount of possible analyses exceeds the number of plausible analyses by far. For forthcoming studies that aim to explore or enrich the North Markian lexicon by spotting novel lexemes, we recommend to operate with cascades 1-3 rather 1-4.

	tokens	forms
total (excl. PUNCT)	5459	1847
coverage (in %)		
cascade 1-3	88.1 (4711/5459)	73.1 (1350/1837)
cascade 1-4	93.4 (5099/5459)	85.1 (1572/1837)

Table 5: Lemmatizer v0.3 coverage over Schwerin (1859), with ('cascade 1-4') and without ('cascade 1-3') word formation, for tokens and distinct forms

4. Related Research

In this paper, we describe the development of a corpus and a computational morphology for a variety of Low German spoken in Germany. To the best of our knowledge, both are the first of their kind for this language, so that related research is somewhat hard to identify.

In terms of language technology, Low German is underresourced, and in particular, there are no corpora available from which NLP tools could be trained that would be capable to process written Low German in German-based orthographies. The corpus Sprachvariation in Norddeutschland (Elmentaler et al., 2015, SiN) features morphosyntactic annotation, and comes with an orthography based on the Sass dictionary of North Low Saxon (Martens, 2004), but is primarily concerned with code-switching between Low German and German regiolects, not with Low German per se. The Zwirner-Korpus (Weber, 2012) contains transcripts of Low German recordings with part-of-speech annotation, but does not present Low German text, but its word-to-word translation into High German. It is thus not usable for the development of NLP tools for Low German. The Low German dataset of the Universal Dependencies (Siewert and Rueter, 2024) is a 20.000 token corpus with syntactic annotations; it is, however, normalized against an innovative – albeit somewhat artificial – interdialectal orthography that deviates substantially from established orthographical traditions based on either High German or Dutch and is thus not widely used beyond the Dutch Low Saxon Wikipedia for which it has been developed. In particular, the corpus cannot be directly used to train NLP tools for the processing of conventional written Low German. Corpora for exilant varieties of Low German have been reported in the literature, e.g., Cox (2010) for Plautdietsch (Canada) and Beilke (2018) for Pomerano (Brazil), but these are not publicly available. Other corpora reported in the literature do not provide naturally occuring text, but elicited language, only, e.g., Misganaw and Roller (2020) and Kaufmann et al. (2023).

In principle, parallel corpora could be used to bootstrap, project or adapt NLP tools even without native annotations, and, indeed, Low German has a sufficiently rich literature to include both literary works translated into even several dialects of the language (the Bible, of course, but also, say, the Little Prince, de Saint Exupéry and Lübben, 2018; de Saint-Exupéry and Demming, 2020, for example), as well as translations from Low German into other languages (e.g., Reuter, 1878, 1888) – but no substantial parallel corpora of Low German are known to exist. A notable exception in this regard is Plautdietsch (Mennonite Low German, ISO 639-3 pdt) for which several Bible translations are available in digital form and from which parallel corpora could be constructed easily – however, Plautdietsch is based on a now-extinct variety formerly spoken around modern Kaliningrad which was characterized by numerous traits not shared with Low German as spoken in modern Germany, and after two centuries of independent development in modern Ukraine and, later, in an international diaspora, both speaker communities have substantially diverged in culture and language.

For historical Low German, the situation is actually much better, as there are several morphosyntactically and syntactically annotated corpora of Old Saxon and Middle Low German (Donhauser et al., 2015; Walkden, 2016; Barteld et al., 2017; Booth et al., 2020), some of which have been semi-automatically annotated with tools similar to what we propose for modern Low German. However, the literary tradition represented by these corpora ended in the 16th c., when Low German as a written language was replaced by modern national languages, esp., German and Dutch. The Low German dialects diversified substantially in the following centuries, and developed novel grammatical traits. This includes, for example, the loss of most unvoiced vowels (apocopy and syncopy) which is characteristic for most modern Low German varieties, and which coincided with a massive restructuring of grammar and syntax. In consequence, the practical relevance of Middle Low German or even older varieties to the development of NLP tools for modern Low German is severely limited.

5. Summary and Outlook

We described the creation of a corpus of North Markian (NMk), based on literary sources from the 19th c., and covering the main regional variants of NMk in terms of lexicon and morphology. Throughout the process, we aimed to minimize manual annotation or disambiguation, but focus on automated and semi-automated methods, instead. We thus relied on automated OCR, without post-correction. Providing only a small seed

dictionary for possible parts of speech for the top 1000 words, and by bootstrapping transition probabilities from corpora of related language varieties, we were able to provide automated POS annotations without actual training data.

This formed the basis for developing a finite state morphology in multiple stages. Without having a lemma list for the language variety under consideration here, we initially relied on handcrafted rules for spelling normalization for each source and for two dictionaries of other Low German dialects. Adopting information about inflectional morphology in a dialect from the interference area with the neighboring MWP variety, this was then complemented with a morphological analyser in order to analyse morphosyntactic features and to provide lemmatization against the MWP, resp. NLS dictionaries. In subsequent refinement steps, we replaced these dictionaries and we developed a North Markian lemma list, a full-form dictionary, and we also revised and extended the morphology to account for all authors (and regions) represented by our corpus. A major result is that the resulting lemmatizer is capable to process unseen North Markian text in German-based orthographies (as demonstrated for Fritz Schwerin). In particular, this includes other texts currently still under copyright which may be more representative for the spoken language than 19th c. poetry, and once these become available, they can be processed with the tools presented here. Also, to some extent other varieties of Low German may be processible with our lemmatizer, if they also exhibit apocopy (loss of unstressed Middle Low German vowels), German orthography and a tendency to avoid diphthongization of Middle Low German ê and ô, e.g., certain varieties of North Low Saxon.

For North Markian, we could confirm and adjust the morphology of Mackel (1905) to account for North Markian beyond the Prignitz region. We created the first lexical resource that represents North Markian directly and in its full extent, as neither of the regional dictionaries (Kettmann, 2004; Stellmacher, 1996) that cover the North Markian dialect region provides North Markian lemmas (but only East Phalian, Central Markian and Central/High German), North Markian appears in the attestations, but without being identified as such and not as structured information. We have to admit, though, that this lexical resource is not a dictionary proper, in that it provides lemma and inflected forms along with their analyses, but without glosses and definitions. At a later point, such information may be drawn from the dictionaries it is largely based on, i.e., Danneil (1859), Müller (1904) and Buck (2007-2024). By normalizing both MWP and NLS dictionaries against North Markian, we also provided a first approach for the interdialectal transliteration of Low German dialects. This ties in very closely with the recent work of Chiarcos et al. (2025a) were we used the same technology to bootstrap a cross-dialectal linking. Both approaches are, however, limited to dictionary linking at the moment. For full-fledged transliteration between Low German *text*, inflectional morphology needs to be accounted for, either by rule-based systems such as presented here, or by neural methods – for which, however, we previously lacked training data. Main results are the lemmatizer and a dictionary that can fill these gaps, along with the corpus; but also that we are able to assess the effort needed for the development of these components in approximate person days (d) and weeks (w):

OCR + text extraction (per source)	∼2 d
HMM POS tagging	∼2 w
Lemmatizer v0.1	∼3 w
Lemmatizer + Dictionary v0.2	∼5 w
Lemmatizer + Dictionary v0.3	∼20 w

Most effort went into curating and verifying the lemma list and full-form dictionaries, but this is an important insight, because for varieties of Low German for which machine-readable lemma lists may already be in existence, development effort is substantially reduced. However, our assessment is conservative in that it does not include any curation of the data itself. We deliberately decided to not correct OCR errors in the source texts, but rather, to deal with that in normalization.

Beyond these findings, the estimated development costs, the bootstrapping workflow, and the resources now available for North Markian open new avenues for research for Low German. The full-form dictionary can serve as a basis for transfer learning, for example by generating synthetic training data for syntactic annotation. More importantly, the data can provide a silver standard that enable experimentation with less labor-intensive approaches: Researchers can experiment with *selected subsets* of our annotated data and compare their performance systematically. This may support workflows that combine minimal, carefully designed training data with lightweight neural methods to achieve comparable results.

The North Markian corpus is released as open source data under CC BY-SA 4.0 and available under https://github.com/nds-spraakverarbeiden/nmk-corpus/, along with Lemmatizer v0.3, Dictionary v0.3 and accompanying scripts published under the Apache License Version 2.0. The corpus can also be accessed online under http://corpora.philhist.uni-augsburg.de.

6. Limitations and Ethical Considerations

As far as ethics are concerned, our material reflects political and social biases of the 19th c., and these are maintained in the data. Dörr (1888) is an anticapitalist novella about rural resistance against financial manipulations. In doing so, it uses common literary tropes of the 19th c., and as part of that, it also builds on the contemporary stigmatization of Jews. Although the book is not antisemitic (in parts, it is actually sympathetic with the Jewish banker who is one of the main antagonists), but (the German version of) its title *Göderschlächter*, meaning 'butcher of estates' has been promoted as an anti-Jewish slogan by the later Nazi regime. Jung (1855) deals with the 1848 revolutionary movement in Prussia, but from a critical, monarchistic, and, occasionally, anti-democratic perspective. Keller (1871) and Keller (1872) deal with the contemporary German-French war and feature nationalistic and anti-French themes.

It may seem a technical limitation that we did not experiment with LLMs as a baseline. As Low German literature is relatively extensive, at least prominent authors such as Fritz Reuter (Mecklenburgian) and the Low German Wikipedias have been present in the training data of multilingual LLMs. Nevertheless, earlier attempts to use LLMs for Low German have been reported as unsuccessful[4] or remain at an embryonic state of development.[5] As for our own work in preparation of the study reported in this paper, conducted a number of experiments with GPT-4o and focused on Reuter and its transliteration to the vernacular of Klaus Groth (also a prominent author whose North Low Saxon dialect has phonological features similar to North Markian), and a randomly selected set of lemmas from both dictionaries. Initially, our prompts were tailored towards creating SFST transducers, but as GPT-4o repeatedly returned FOMA syntax, we eventually asked for FOMA. These resulting transducers effectively performed 1:1 mappings and the removal of diacritics. To put more focus on the mapping task itself, we then changed the prompt to produce a JSON dictionary with character replacements. Again, these were effectively 1:1 replacements and diacritic removal. Without any promising result, we abandoned these experiments after two working days. For comparison, writing the Reuter normalization by hand took 3 hours. It is unsurprising that LLMs largely fail at this task because even though they certainly have seen some Low German data as part of their training, certainly including writings of Reuter and Groth,

they have probably not encountered North Markian (and if so, only in poor quality OCR) and they seem to have difficulties to generalize over the multitude of orthographies in a way that allows for transliteration. This may be different when asking the system to transliterate directly, but this clearly is an unjustifiable waste of energy in comparison to the little human effort it takes to come up with a mapping table.

An alternative future direction for LLM-based techniques may include encapsulating FSTs or a cross-dialectal dictionary lookup as agentic components, so that the language capacity (implicitly normalized towards the majority dialects as represented in the training data) of the model is separated from its transliteration capabilities (for which it lacks training data). With FSTs for normalization and inflection, fundamental components for such a system are provided along with this paper.

In fact, this is a direction that may address another limitation: This release does not yet include the contextual disambiguation of morphological analyses. Initial experiments to align the phonological representation of the corpus with the Low Saxon LSDC Dataset at Universal Dependencies (Siewert and Rueter, 2024) have been unsuccessful. The primary reason is that the orthography used in the LSDC dataset requires to write Middle Low German unvoiced vowels that are systematically lost in North Markian. For aligning the transliterations, these thus need to be restored, by means of a rule that allows to insert unvoiced vowels effectively *anywhere* (when normalizing from North Markian to the Nysassiske Skryvwyse adopted by the LSDC dataset) or to arbitrarily drop e (when normalizing from LSDC to our internal normalization). But inserting and dropping vowels also comes with different assimilation processes affecting the neighbouring consonants, whose technical implementation is laboursome. It seems more tractable to focus on transliteration between Low German varieties with and without apocope, and albeit the latter seem to represent the majority of Low German dialects in Germany and abroad (excluding West Phalian, but including Mennonite Low German, or, Plautdietsch and the Pomerano variety of Brazil; but this is less frequently found in Low German varieties in the Netherlands), no

[4] https://www.ndr.de/kultur/norddeutsche_sprache/niederdeutsch/Pepper-Blog-34-Neue-wissenschaftliche-Wege,pepperblog180.html, accessed March 6, 2026

[5] https://www.ndr.de/kultur/norddeutsche_sprache/plattdeutsch/archiv-luenki-blog-der-kleine-ki-spatz-lernt-plattdeutsch,luenkiblog-104.html, accessed March 6, 2026; also see https://digital-zentral.de/kuenstliche-intelligenz-auf-plattdeutsch-eine-sprachliche-herausforderung/ (accessed March 6, 2026) for a general overview over artificial intelligence in application to Low German.

resources for such varieties seem to exist.

Another limitation is the exclusive reliance on automated digitizing of historical texts, which may affect the accuracy and usability of the final corpus. For the data at hand, this is partially compensated by a publication via TEITOK (Janssen, 2016), which allows privileged users to apply corrections directly to the corpus. As we aim to engage with the aging speaker community, which seems to be interested in the documentation and preservation of their language in the digital sphere, and the use of digital tools to facilitate its use and transmission, a longer perspective may be seen in crowdsourcing error corrections. The texts used as a basis here are of cultural significance for their respective region, so that engaging with speaker community may pave the way for a larger-scale crowdsourcing effort.

7. Bibliographical References

Fabian Barteld, Katharina Dreessen, Sarah Ihden, and Ingrid Schröder. 2017. Das Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200–1650) – Korpusdesign, Korpuserstellung und Korpusnutzung. *Mitteilungen des Deutschen Germanistenverbandes*, 64(3):226–241.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology: Xerox tools and techniques*. CSLI Publications, Stanford.

Neubiana Silva Veloso Beilke. 2018. Pommersche korpora: um conjunto de corpora dialetais da variedade brasileira do pomerano. *Linguística de Corpus: Perspectivas. Porto Alegre: Instituto de Letras–UFRGS*, pages 365–398.

Hannah Booth, Anne Breitbarth, Aaron Ecay, and Melissa Farasyn. 2020. A Penn-style Treebank of Middle Low German. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC-2020)*, pages 766–775.

Wilhelm Bornemann. 1810. *Plattdeutsche Gedichte*. Decker, Berlin.

Wilhelm Bornemann. 1816. *Plattdeutsche Gedichte. Zweites Bändchen*. Decker, Berlin.

Wilhelm Bornemann. 1868. *Plattdeutsche Gedichte*. Decker, Berlin.

Marcus Buck. 2007-2024. Plattmakers – Das Plattdeutsche Wörterbuch. https://plattmakers.de. Accessed November 9th, 2024.

Christian Chiarcos, Tabea Gröger, and Christian Fäth. 2025a. Putting Low German on the map (of Linguistic Linked Open Data). In *Proceedings of the 5th Conference on Language, Data and Knowledge*, pages 62–75, Naples, Italy. Unior Press.

Christian Chiarcos, Janine Siewert, Tabea Gröger, and Christian Fäth. 2025b. Towards a cross-dialectal dictionary for Low German (Low Saxon). In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Long and Short Papers*, pages 282–294, Hannover, Germany. HsH Applied Academics.

Christopher Cox. 2010. Probabilistic tagging of minority language data: a case study using qtag. *Corpus linguistic applications: current studies, new directions. Amsterdam: Rodopi*, pages 213–231.

Johann Friedrich Danneil. 1859. *Wörterbuch der altmärkisch-plattdeutschen Mundart*. J. D. Schmidt, Salzwedel.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Antoine de Saint-Exupéry and Hannes Demming. 2020. *De lütke Prins: Der kleine Prinz - Münsterländer Platt: Mönsterländsk Plat*. Edition Tintenfaß, Neckarsteinach, Germany.

Antoine de Saint Exupéry and Antje Lübben. 2018. *De lüttje Prinz: Der kleine Prinz - Ostfriesisch Platt: Der kleine Prinz Ostfriesisches Platt*. Edition Tintenfaß, Neckarsteinach, Germany.

Stefanie Dipper. 2015. Annotierte Korpora für die Historische Syntaxforschung: Anwendungsbeispiele anhand des Referenzkorpus Mittelhochdeutsch. *Zeitschrift für germanistische Linguistik*, 43(3).

Karin Donhauser, Jost Gippert, and Rosemarie Lühr. 2015. Das Referenzkorpus Altdeutsch. In Jost Gippert and Ralph Gehrke, editors, *Historical Corpora. Challenges and Perspectives*, pages 35–49. Narr, Tübingen.

Julius Dörr. 1888. *De Göderschlächter. Dörpgeschicht ut de Uckermark*. Hesse & Becker, Bad Freienwalde.

Michael Elmentaler, Joachim Gessinger, Jens Lanwer, Peter Rosenberg, Ingrid Schröder, and Jan Wirrer. 2015. Sprachvariation in Norddeutschland (SiN). In *Regionale Variation des Deutschen. Projekte und Perspektiven*, pages 397–424. Mouton De Gruyter, Berlin and Boston.

Rudolf Hill. 1868. *Lütte Schnurren. Plattdeutsche Gedichte*. Vincent, Prenzlau.

Maarten Janssen. 2016. TEITOK: Text-faithful annotated corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4037–4043.

Gustav Jung. 1855. *Gedichte in plattdeutscher Mundart*. Wohlgemuth, Berlin.

Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 2017. Transkribus - A service platform for transcription, recognition and retrieval of historical documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR-2017)*, volume 4, pages 19–24. IEEE.

Göz Kaufmann, Jan Gorisch, and Thomas Schmidt. 2023. Das mend-korpus im archiv für gesprochenes deutsch: Entstehung, möglichkeiten, grenzen. In *Deutsche und weitere germanische Sprachminderheiten in Lateinamerika. Grundlagen, Methoden, Fallstudien*, pages 103–147. Peter Lang.

Ernst Keller. 1871. *De Pommersche Landwehrmann Crischon in'n französischen Krieg*. Dannenberg, Stettin.

Ernst Keller. 1872. *Crischon Ballermann, Garde-Landwehrmann von't Stettiner Batteljon*. Dannenberg, Stettin.

Ernst Keller. 1877. *Der Angerhof. Volksstück mit Gesang in 3 Acten*. Selbstverlag, Pyritz.

Gerhard Kettmann. 2004. Das Mittelelbische Wörterbuch – Die problemreiche geschichte eines notwendigen forschungsprojektes. *Mitteldeutsches Jahrbuch für Kultur und Geschichte*, 11:27–34.

Emil Mackel. 1905. *Die Mundart der Prignitz*. Soltau.

Peter Martens. 2004. (Review of) Der neue SASS. Plattdeutsches Wörterbuch. Plattdeutsch - Hochdeutsch. Hochdeutsch - Plattdeutsch. Plattdeutsche Rechtschreibung. *Zeitschrift für Dialektologie und Linguistik*, 71(1):113–115.

Aynalem Tesfaye Misganaw and Sabine Roller. 2020. Plattform: Parallel spoken corpus of middle west german dialects with web-based interface. In *Machine Learning, Optimization, and Data Science*, pages 494–503, Cham. Springer International Publishing.

Carl Friedrich Müller. 1904. *Reuter-Lexikon: der plattdeutsche Sprachschatz in Fritz Reuters Schriften*. Hesse & Becker.

Hermann Pfaff. 1898. *Die Vocale des mittelpommerschen Dialects*. A. Straube.

Fritz Reuter. 1878. *An old story of my farming days*. Bernhard Tauchnitz, Leipzig. Translated by M.W. MacDowell.

Fritz Reuter. 1888. *Matkustus Belgiaan*. Turku. Translated by Werner Söderström.

Helmut Schmid. 2005. A programming language for finite state transducers. In *International Workshop on Finite-State Methods and Natural Language Processing*, pages 308–309. Springer.

Fritz Schwerin. 1859. *Vöggel-Sproak un Snack, oder: Was die Vögel klein und gross im Frühjahr in der Altmark singen und sagen*. Eyraud, Neuhaldensleben.

Erich Seelman. 1908. *Die Mundart von Prenden (Kreis Nieder-Barnim)*. Ph.D. thesis, Universität Breslau, Norden.

Janine Siewert and Jack Rueter. 2024. The Low Saxon LSDC dataset at Universal Dependencies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15976–15981, Torino, Italia. ELRA and ICCL.

Dieter Stellmacher. 1996. Brandenburg-Berlinisches Wörterbuch. Begründet und angelegt von Anneliese Bretschneider unter Einschluss der Sammlungen von Hermann Teuchert, fortgesetzt von Gerhard Ising, bearbeitet unter der Leitung von Joachim Wiese. Bd. 3, Lfg. 1-10. *Zeitschrift für Dialektologie und Linguistik*, 63(3):339.

Dieter Stellmacher. 2017. *Niederdeutsch: Formen und Forschungen*. Walter de Gruyter, Berlin.

Hermann Teuchert. 1944. *Die Sprachreste der niederländischen Siedlungen des 12. Jahrhunderts*. Karl Wachholtz Verlag, Neumünster.

George Walkden. 2016. The HeliPaD: A parsed corpus of Old Saxon. *International Journal of Corpus Linguistics*, 21(4):559–571.

Thilo Weber. 2012. Neue Fragen an alte Daten. Niederdeutsche Syntaxgeographie auf der Grundlage von Zwirner-und DDR-Korpus. In *Niederdeutsche Syntax*, pages 157–180. Olms.

Peter Wiesinger. 1982. Probleme der Dialektgliederung des Deutschen. *Zeitschrift für Dialektologie und Linguistik*, pages 145–168.

A. Appendix

As for the specific features of Low German in comparison to other West Germanic languages, Tab. 6 illustrates characteristics of the respective varieties and the corresponding realizations in Low German. According to a traditional view in historical linguistics, Low German occupied an intermediate position between the historical North Sea Germanic dialects (predecessors of English and Frisian) on the one hand, and continental West Germanic varieties (predecessors of Dutch and High German) on the other in the historical dialect spectrum of West Germanic, and this is still reflected in the modern languages.

Table 7 illustrates the degree of orthographic variation within North Markian as well as deficits of the respective German-based orthographies. Because the different author orthographies preserve different aspects of the underlying phonology, we decided to provide one FST per author, as the same symbol may have a different interpretation in another context. In particular, this was used to develop the full form dictionary. As for Lemmatizer v0.3, this is primarily designed to analyze unseen text, so that a uniform normalization is provided by the join of the author-specific FSTs.

Figure 4 provides a graphical visualization of the iterations of the Lemmatizer, its analytical components and the associated dictionaries. To illustrate the functionalities at different stages of development, we also provide a sample analysis for the clause *ik heff't furts wedder instaken* 'immediately, I put it inside again' from (Dörr, 1888). For this example, Table 8 illustrates the initial results of normalization and of Lemmatizer v0.0, i.e., normalization and lookup in North Low Saxon (NLS, Plattmakers) and Mecklenburgian (MWP, Reuter) dictionaries. Overall, the normalization yields reasonable results, but for any inflected form (as well as for any form not in the dictionaries), the lookup fails, so that here, only 2 of five words can be analyzed.

To this baseline, Lemmatizer v0.1 added inflectional morphology and the handling of clitics as shown in Table 9. Note that it is generally assumed that North Markian (like the neighbouring Mecklenburgian dialect) lost the distinction between accusative and dative, but as this has never been empirically explored, and other neighboring dialects (Central Markian and Eastern Pomeranian) actually preserved aspects of an older systematic distinction between accusative and dative, these are distinguished in the morphology for the moment.

The analysis of Lemmatizer v0.1 is promising, in that almost every word – except for *furts* 'immediately' – is assigned one correct analysis. However, it is still grounded in the Plattmakers and Reuter dictionaries, i.e., varieties that followed independent developmental trajectories for about half a millenium. For Lemmatizer v0.2, the resulting analyses were manually pruned and North Markian lemmas were constructed, with pronounciation reconstructed from the textual evidence and based on the etymological correspondences with Mecklenburgian, North Low Saxon, German, English and Dutch. Lemmatizer v0.2 also added support for word formation, but this was found to overgenerate heavily. Therefore, Lemmatizer v0.3 exposes a subset of Lemmatizer v0.2 analyses only, it suppresses deeper Lemmatizer v0.2 analyses if analyzable forms can be found in the dictionaries or the mere inflectional morphology yields an attested lemma, cf. Table 10. The following analyzes (and their morphosyntactic variants) produced by Lemmatizer v0.2 were suppressed by Lemmatizer v0.3:

ik	`eken/VERB.1.Sg.Ind.Prs`
	(hypothetical verb formed for ek "corner")
heff't	`haft/NOUN.Fem.Pl_uml`
	("captivity", multiple forms)
furts	`fertsen/VERB.1.Sg.Ind.Prt.st.iii`
	("to fart", dialectal High German)
	`furtsen/VERB.1.Sg.Ind.Prs`
	("to fart", High German)
wedder	`wed/NOUN.Fem.Pl_er.Nom`
	("bets", multiple forms)
instaken	`in+steken/VERB.PPast.st`
	(correct, verbal particle split off)

Each of these has multiple morphological analyses, again, not shown here for reasons of space. All of these are possible linguistic analyses, although the hypothetical verb `eken` does not exist, the plural of `wed` "bet" is normally formed with *-en*, not with *-er*, and neither does `haft` "captivity" form a plural by umlaut.

English	Frisian (North Frisian) https://friesisch.net/	Low German (North Low Saxon) https://plattmakers.de	Dutch	High German	linguistic process	approximate date
cheese *church*	*säis, sees* *schörk, sark*	*Kees* *Kark*	*kaas* *kerk*	*Käse* *Kirche*	Anglo-Frisian palatalization	5th -7th c.
make *apple*	*måååge, maage, maaki* *ååpel, ååple, aapel*	*maken* *Appel*	*maken* *appel*	*machen* *Apfel*	High German consonant shift	5th - 8th c.
house /aʊ/ *fine /aɪ/*	*hüs, hüüs /y:/* *fiin /i:/*	*Huus /u:/* *fien /i:/*	*huis /œy/* *fijn /ɛi/*	*Haus /aʊ/* *fein /aɪ/*	Diphthongizations (Great Vowel Shift)	13th - 17th c.

Table 6: Linguistic features of Low German in comparison to other West Germanic languages

		selected words			phrasal expressions	
		“to have” (AUX)	“for this” (ADV)	“to beat” (VERB)	“old women”	“had we it”
phonology	IPA	/hɛvn̩/ ~ /hɛbn̩/ ~ /hɛbm̩/ ~ /hɛmm̩/ ~ /hɛmn̩/	/dɔ:fœe̯/	/ʃlɔ:n/ ~ /slɔ:n/ (/ʃlɔ:gŋ/)	/ɔl fru:n̩s/ (and variants)	/harn̩ vi:t/ (and variants)
	internal	heben	dåfœr	Slån (Slågen)	ol(e) frUns / frugens	haren wI’t
	lemmas	heben	dåfœr	Slån (Slågen)	old frU	heben wI et
orthography	cf. Plattmakers	*hebben*	*dor för*	*slahn (slagen)*	(not applicable to dictionaries)	
	cf. Reuter	*hewwen*	*dorför*	*slahn (slagen)*		
	Danneil	*hemm, hemm’n*	*daoväör*	*slaon*		
	Bornemann	*hemm, hem, hem’, hem’n, hebben*	*doavör, davör*	*schloan*	*olle Fru’ns*	*har’n wie’t, harren wie’t*
	Jung	*hemm, hemm’*	*doavöär*	*schloan*	*olle Fru’ns*	*haarn wie’t*
	Dörr	*hebben*	*davær*	*(schlagen)*	*olle (oll) Frugens*	*hadden wi’t*
	Hill	*hebben, hebb’n*	*doavöar, doaföar*	*schloan*	*oll Fru (Sg.)*	*hadd’n wie’t*
	Keller	*hebben, hebb’n, heww’n, hewwen*	*daför, doaför doavör*	*(schloagen)*	*oll (olle) Frunslüd / Frug’nslüd*	*harnw’t, harrn wi’t*

Table 7: Selected words and (in parts, constructed) phrases in (normalized) phonological representation and original orthographies of authors, the dictionary of the Altmark region by Danneil and two dictionaries of North Low Saxon (Plattmakers) and Mecklenburgian (Reuter)

		candidate normalizations			Lemmatizer v0.0			
					Plattmakers		Reuter	
ik	I	***ik***			***ik/PRON*** **(I)**		***ik/PRON*** **(I)**	***ick/PRON*** **(I)**
heff’t	have it	***hef’t***			no analysis			
furts	immediately	***furts***	***fOrts***	***fUrts***	no analysis			
wedder	again	***ved’der***	***wed’der***		***wedder/ADV*** (again)	***Wedder/NOUN*** (wheather)	***wedder/ADV*** (again)	***Weder/NOUN*** (wheather)
instaken	put.inside	***in’sta’ken***	***in’stå’ken***	***in’stä’ken***	no analysis			
		$FST_{a \to norm}$			$FST_{a \to norm} \circ FST^{-1}_{d \to norm} \circ FST_{d}$			

Table 8: Normalization and dictionary lookup (‘Lemmatizer v0.0’) for an example clause

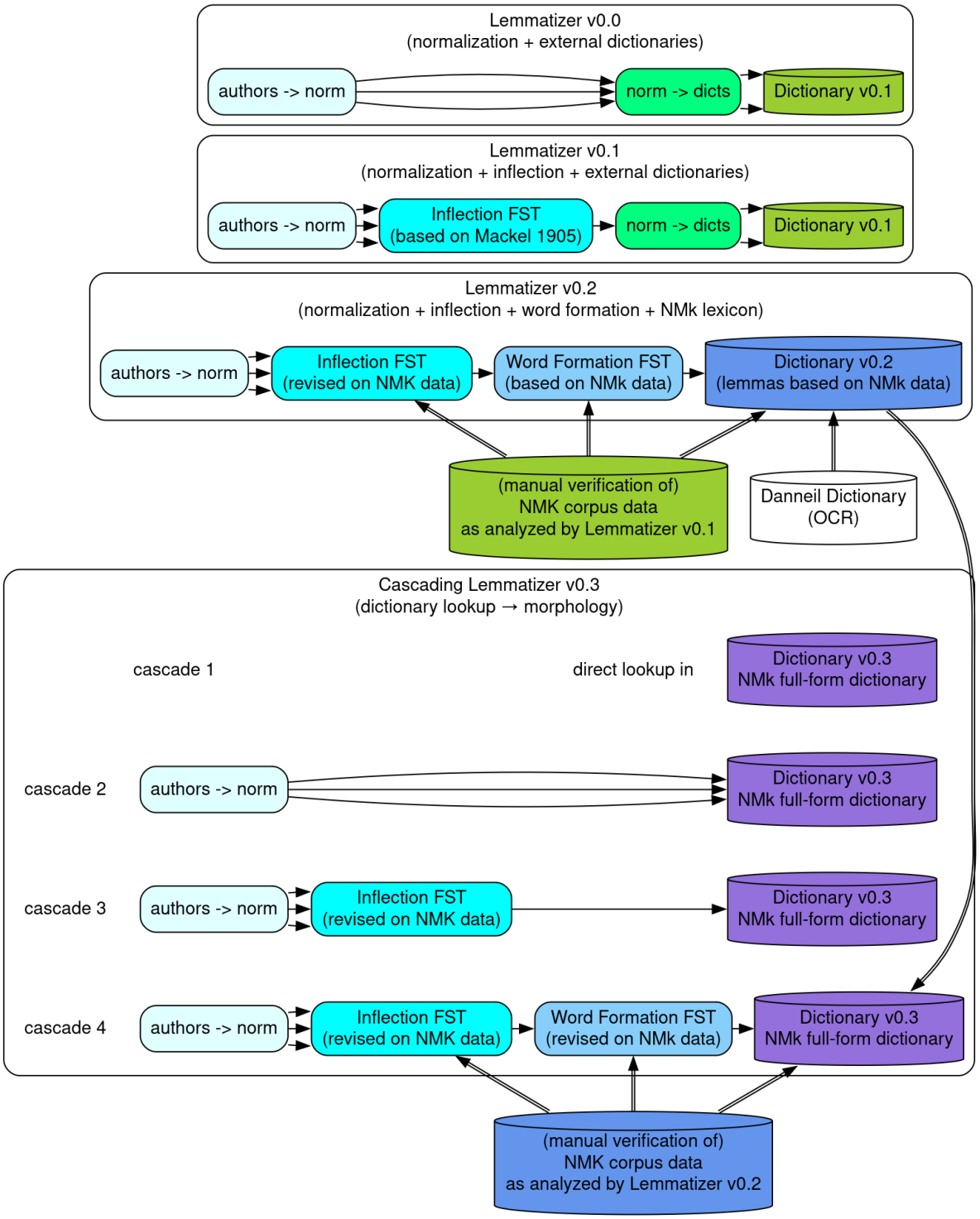

Figure 4: Schematical overview illustrating the evolution of the Lemmatizer, its analytical components, the associated dictionaries and the flow of information between them

WORD	ANALYSES AND COMMENTS	
ik	ik/PRON.1.Sg.Nom.nds	(*"I", correct*)
	ik/PRON.1.Sg.Nom.mvp	
	ick/PRON.1.Sg.Nom.mvp	
heff't	hebben/AUX.1.Sg.Ind.Prs.nds+et/PRON.3.Sg.Neut.Nom.Sg.clit.nds	
	...+et/PRON.3.Sg.Neut.Acc.Sg.clit.nds	
	(*"have it", correct*) ...Dat...	
	...+de/DET.Def.Nom.Sg.Masc.clit.nds	
	...Dat.Sg.Masc...	
	...Nom.Sg.Fem...	
	...Acc.Sg.Fem...	
	...Dat.Sg.Fem...	
	...Nom.Pl...	
	...Acc.Pl...	
	...Dat.Pl...	
	...+dat/DET.Def.Nom.Sg.Neut.clit.nds	
	...Acc.Sg.Neut...	
	...Dat.Sg.Neut...	
	...+dat/PRON.Dem.Neut.Sg.Nom.clit.nds	
	...	
	hebben/AUX.2.Pl.Imp.nds	
	hewwen/AUX.1.Sg.Ind.Prs.mvp+...	
	hewwen/AUX.3.Sg.Ind.Prs.mvp+...	
	hewwen/AUX.2.Pl.Imp.mvp	
	hewen/VERB.1.Sg.Ind.Prs.mvp+...	(*"to bear, carry, lift up"*)
	hewen/VERB...	
furts	(*"immediately", not found, Plattmakers only has the cognate adverb furtsen*)	
wedder	wedder/ADV.nds	(*"again", correct*)
	wedder/ADV.mvp	
	Wedder/NOUN.Nom.Sg.nds	(*"wheather"*)
	Wedder/NOUN.Acc.Sg.nds	(*no gender information, yet*)
	Wedder/NOUN.Dat.Sg.nds	
	Wedder/NOUN.Nom.Pl.nds	
	...	
	Weder/NOUN.Nom.Sg.mvp	(*"wheather"*)
	...	
instaken	insteken/VERB._.Pl.Ind.Prt.st.nds	(*"put.inside"; correct, but no*
	insteken/VERB.PPast.st.nds	*ablaut/inflection class, yet*)
	insteken/VERB._.Pl.Ind.Prt.st.mvp	
	insteken/VERB.PPast.st.mvp	

Table 9: Lemmatizer v0.1 output (inflectional morphology); against Plattmakers [nds] and Reuter [mvp]

WORD	NORM	LEMMA/MORPH	
ik	ik	ik/PRON.1.Nom.Sg	
heff't	heft	heben/AUX.1.Sg.Ind.Prs+PRON.3.Sg.Neut.Acc.clit	
furts	furts	fOrts/ADV	(*this is correct, related to Italian forza*)
		furts/NOUN.Masc.Nom	(*these are incorrect, but*
		furts/NOUN.Masc.Acc	*possible, for a High*
		furts/NOUN.Masc.Dat	*German loan for "fart"*)
wedder	weder	weder/ADV	
instaken	inståken	insteken/VERB.PPast.st	

Table 10: Lemmatizer v0.3 output; against North Markian lemma list and full-form dictionary