Towards a Cross-Dialectal Dictionary for Low German (Low Saxon)

Christian Chiarcos* **and Janine Siewert**** **and Tabea Gröger*** **and Christian Fäth***
*Applied Computational Linguistics (ACoLi), University of Augsburg, Germany
**Department of Digital Humanities, University of Helsinki, Finland
first_name.second_name@uni-a.de / ...@helsinki.fi

Abstract

We describe and evaluate a methodology for the automated creation of an interdialectal lexical resource (a "digital Rosetta stone") for major dialects of Low German (Low Saxon) spoken in Germany, based on the interlinking of lexical resources freely accessible over the web. The resulting dataset is provided both in human-readable form and as a lexical knowledge graph compliant with Linked Data standards, as content from dictionaries under copyright or non-derivative licenses could only be linked from but not included in our release data.

1 Motivation

The 'digital fitness' of languages (Soria et al., 2016) is often measured based on parameters such as the existence of resources and tools in relation to the size of the speaker community. However, depending on the *degree of internal diversity* of the language, the usefulness of, say, machine translation or dictionaries, varies greatly for the different subgroups of the community. *Low German* or *Low Saxon* is a language with such a high degree of internal variation.

As a West Germanic language, Low German is related to (High) German, Dutch and Frisian, and spoken in northern Germany, parts of the Netherlands and by several emigrant communities (Fig. 1 and Fig. 2). Low German has a literary tradition of more than 1000 years (Sievers, 1875) and was a lingua franca around the Baltic Sea during the late Middle Ages, but during the early modern period, it was replaced as a written language by the emerging national languages in Germany and the Netherlands. Since the 1990s, it has been recognized as a regional language in both countries under the European Charter for Regional or Minority Languages (ECRML), but has been and still is subject to substantial pressure from the respective national languages (Schwenk, 2017; Adler and Beyer, 2017). Low German enjoys cultural and regional recognition, but revitalization efforts have so far remained relatively unambitious compared with other European minority languages such as Catalan, Welsh and North Sámi (Šatava, 2019). The integration of Low German into digital spaces and modern media still requires great work for it to retain and regain the status of a thriving and fully functioning language (Blaschke et al., 2023).

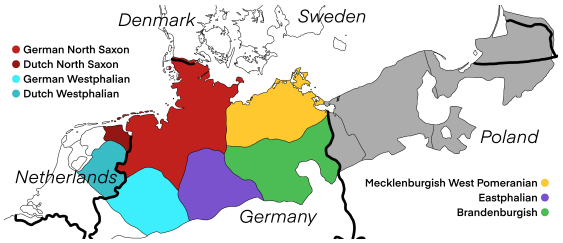

Figure 1: The major dialect groups of Low German (Low Saxon). Marked in grey are the eastern dialects which were spoken in these areas until the end of WWII.

In particular, the current state of linguistic and orthographic fragmentation poses problems to the construction of many basic resources, either for NLP or didactic purposes (Reershemius, 2010; Ehlers, 2021; Bieberstedt, 2021). As it lacks an interregional standard, Low German exhibits not only dialectal variation in phonology and grammar, but speakers also vary in their ways of spelling the language due to which speakers of the same variety cannot automatically be assumed to have the same orthographic preferences (Teuchert, 1914; Martens, 1979, 2002; Weber and Schürmann, 2018).

Without parallel corpora and a complete lack of machine-readable dictionaries, we argue here that what is needed as a minimal requirement for the development of an effective NLP support for Low German – say, standard tools such as spell-checking, search, machine translation, chatbot technology, but also transliteration of texts from other Low German dialects –, is a digital Rosetta stone

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 1: Long and Short Papers, pages 282–294, 2025

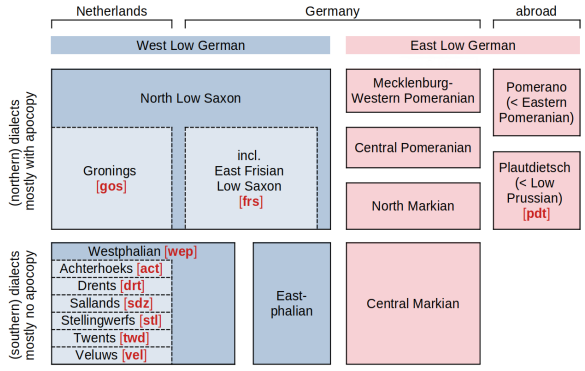

Figure 2: Major dialects of Low German (ISO 639-2 nds), with regional ISO 639-3 codes in red square brackets.

that allows us to leverage materials from different language varieties and to access them in a uniform way. However, instead of creating an artificial standard variety, we focus on technologies to complement dialect specific resources, by creating links between regional dictionaries, and by providing a mapping routine to identify formally corresponding words in different dialects of Low German.

The resulting dataset is provided both in human-readable form and as a lexical knowledge graph compliant with Linked Data standards,[1] as content from dictionaries under copyright or non-derivative licenses could only be linked from but not included in our release. Data, linking and conversion scripts are available from the *NDS Spraakverarbeiden* organization at GitHub.[2]

2 Data Sources and Data Modelling

We provide an interdialectal lexical resource for major dialects of Low German in Germany by interlinking lexical resources freely accessible over the web. Our contribution is two-fold: We describe a methodology for linking and create a cross-dialectal lexical knowledge graph for Low German that may serve as a basis for the subsequent development and application of statistical or neural methods of machine learning.

[1]For a general introduction into Linked Data and lexical knowledge graphs in and for language technology, technologies and use cases, see Cimiano et al. (2020).

[2]For data and description, see https://nds-spraakverarbeiden.github.io/linked-nds-dictionaries/, our source repository is under https://github.com/nds-spraakverarbeiden/linked-nds-dictionaries.

2.1 Dictionaries and Dialects

We operate with 6 digital dictionaries of major varieties of Low German in Germany:

Sass (Thies, 2025, North Saxon)
Plattmakers (Buck, 2007-2024, North Saxon)
WöWö (Neuber, 2001, North Saxon/Ditmarschen), see Sect. 2.2
PlattWB (Brückmann, 2025, North Saxon/East Frisian)
Reuter (Hansen, 2025, East Low German/Mecklenburgian), digital edition of Müller (1904)
WWB (Niebaum et al., 1969-2021, Westphalian)

In addition, we use a North Saxon/Holsteinian glossary with manually linked WöWö lemmas for evaluation (Sect. 4.1).

Of the aforementioned, more substantial dictionaries, only WöWö (published as 'Frie Woor') seems to allow the creation of local copies from which partial structured information (i.e., glosses) can be extracted. It is thus used as a basis for the lexical knowledge graph we aim to provide. Beyond WöWö lemmas and definitions, we only include URIs and lemma forms for external dictionary content that can be linked with WöWö. In order to avoid copyright infringement, we did not use the actual content of the external dictionaries, but provide a purely form-based linking, only. The legal situation also has an impact on the technologies we can use for publishing the results of our linking procedure: As the original data resides with its providers, scattered throughout Germany, we resort to Linked Open Data (Bizer et al., 2009, LOD), resp., the underlying RDF technologies, as this allows to release our data in a way that integrates information from remote hosts and abstracts away from their individual choice of tools, formats and technologies to provide it. A minimal requirement in that regard is that they allow to address their lexical entries by means of a URI.

2.2 Converting the Wöhrner Wöör (WöWö)

In recent years, a lot of efforts have been made to establish an interconnected pool of language resources under an open license as Linguistic Linked Open Data (LLOD).[3] A major corner stone to guarantee interoperability is the consistent use of RDF ontologies and data models, and we consequently

[3]https://linguistic-lod.org

employ OntoLex,[4] the de-facto standard for publishing lexical data as LLOD, for data modelling.

The OntoLex core model establishes classes for key elements of any lexicon, each identified with a Uniform Resource Identifier (URI): the **LexicalEntry** represents an entry in a dictionary, a **Form** is a surface form or phonetic realization, and the `LexicalSense` represents the word sense. In addition to these, various relations can be specified, e.g., a sense can refer to other senses or to an external resource. Additional modules provide further descriptors for modeling additional concepts or relations, and we employ the `vartrans` module for representing cross-dialectal links between multiple entries in different dialects of Low German.

The core of the lexical knowledge graph stems from the *Wöhrner Wöör* (WöWö) dictionary of the Dithmarschen variety of North Saxon, published in print in 2001 by Peter Neuber, followed by an extended and revised digital version 2019 as PDF and MS Office documents that now comprises a total of 21,012 German and 26,702 Low German lexical entries.[5] Aimed at human readers, the WöWö lacks structured machine-readable representations. In accordance with established best practices for bilingual dictionaries (Gracia and Vila-Suero, 2015) and in order to facilitate its linking with other lexical datasets, we thus converted it into an OntoLex RDF graph as summarized in Fig. 4. As shown in Figure 3, a variety of colors, fonts, and fonts sizes are used to encode different pieces of information which causes a fragmentation of the underlying text information and made the conversion of the *Wöhrner Wöör* challenging. These different fragments were merged by the extractor as part of a multi-stage conversion process implemented in Python. For details of the data modelling process and the conversion process, see Chiarcos et al. (2025). For reasons of copyright, the WöWö remains our only source of sense information. For the other dictionaries mentioned above, we only make use of information about geographical variant forms (and their respective lemma URL).

2.3 Linking External Dictionaries

Aside from WöWö, the other Low German dictionaries considered here are accessible online, with URIs identifying the respective lemma, and we use only *this information* (the existence of a lemma and the assignment of a particular URL) to extend the WöWö core graph with an index for these in RDF. We assume that this information does not meet the threshold of originality legally required for copyright to apply (Margoni, 2016).

The dictionaries we work with are isolated from any other content available on the web. Yet, this does not mean that they do not contain links. In fact, *several* of the platforms from which the aforementioned dictionaries are drawn have been *designed* to provide inter-dialectal links, resp., links between different dictionaries,[6] but they only provide links *within* the respective ecosystem, whereas we pursue an open, extensible approach to integrate *any* piece of information accessible on the web.

Creating an LOD index for a dictionary requires to retrieve its complete content, to extract lemma forms and lemma URL and to store these in a TSV file. By means of the linking discussed in Sect. 3, we extend this file with the lemma form in WöWö, the WöWö URL (the actual link), and a confidence score (for pruning and verification, see Sect. 3.2). As only lexical forms are extracted from the external dictionaries, linking is solely grounded on agreement on the level of *forms*, without tackling the dimension of meaning. Figure 5 illustrates a case of 2:1 linking for a dictionary from a Low German dialect from the Netherlands (Twents), where the entries *Oal* 'slimely person' and *oal* 'eel' are linked to WöWö *Ool* 'eel', using `https://twentswoordenboek.nl` as a basis.

3 Lexical Linking by Formal Agreement

There is no single standard variety or standard orthography for Modern Low German, and the existing dialects differ in their phonology. In general, spellings of northern Low German in Germany closely follow the Standard German orthography. These are defective in the sense that certain phoneme differences are not systematically represented. In particular, this pertains to the three- or four-fold differentiation between several series of historically long *ê*, *ô*, and *ô̂* and their short and lengthened counterparts, whose writing cannot be

[4] `https://www.w3.org/2016/05/ontolex/`

[5] `https://ditschiplatt.de/woehrner-woeoer/`

[6] The WWB is published as part of the Trier Wörterbuchnetz (`https://woerterbuchnetz.de/`), which *can* provide HTML hyperlinks between different dictionaries, but it has been the only Low German dictionary on that platform. In late April 2025, a Mecklenburgian dictionary was added, but too late to be addressed in this paper. The Reuter dictionary is part of the Digitales Wörterbuch Niederdeutsch (DWN) (`https://www.niederdeutsche-literatur.de/dwn/`), along with three other Low German dictionaries whose print sources, however, are still under copyright.

Figure 3: Excerpt of the *Wöhrner Wöör* dictionary.

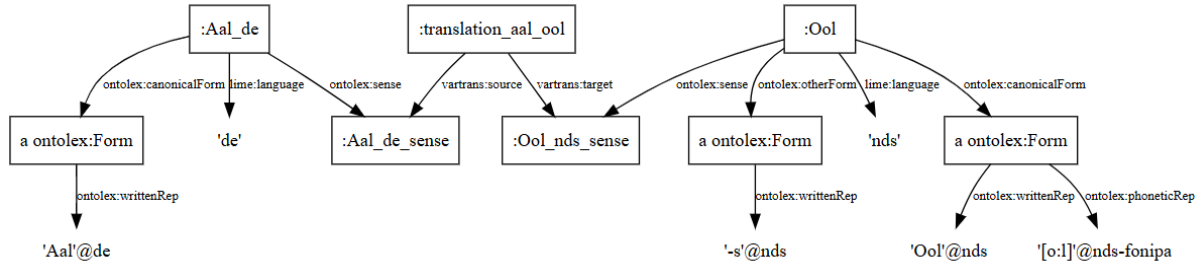

Figure 4: RDF conversion of WöWö lemma *Ool* and its German headword *Aal*, cf. second line of Fig. 3.

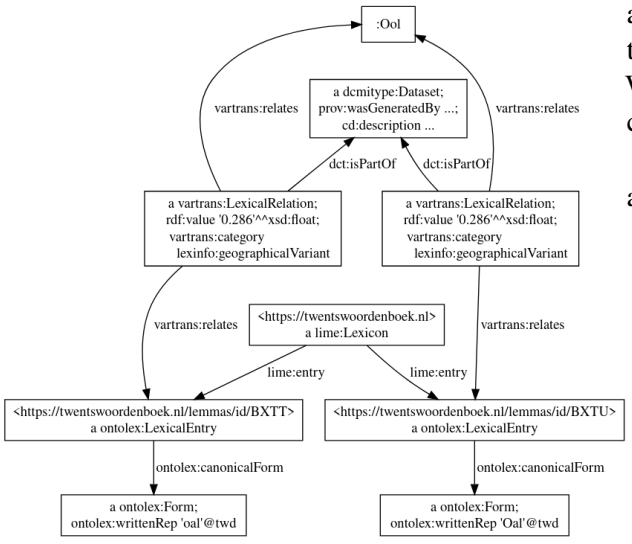

Figure 5: Reified `lexinfo:geographicalVariant` links between WöWö *Ool* ‘eal’ and Twents dictionary.

grounded on the ‘intuitive’ application of Standard German spelling conventions designed to express the two-fold distinction between long and short vowels only. Most southern Low German dialects are phonologically even richer and employ a multitude of custom orthographies. The Westphalian WWB dictionary uses a scientific notation for reconstructed interdialectal forms.

Phonological differences exhibited by the dialects of Low German include, for example

- **mergers**: Historically, Low German had four long *ê* phonemes and two long *ô* phonemes including their umlauted forms, while the modern dialects have merged some of these: While Reuter distinguishes $\hat{o}_1$ in *Kauken* ‘cake’ and $\hat{o}_2$ in *Bom* ‘tree’ , WöWö *Kōken* and *Bōōm* suggests a merger.

- **apocope and syncope**: unstressed Middle Low German short vowels are mostly lost in the north, and retained in the south.

- **lengthening and shortening of vowels**: applied in all dialects, but with different results. These were triggered by Middle Low German

syllable structure, which, however, became obfuscated by apocope and syncope.

- **assimilation** of/with post-vocalic consonants, esp. *r*, led to regionally different results, e.g., for *Barg* 'mountain' (WöWö) alongside *Bo^{a}rg* (WWB).

Historical phonology and the characteristics of the respective orthographies are well understood, so that rules to map from one orthography to another can be relatively easily implemented on the basis of standard technology such as Finite State Transducers (FSTs). Because this involves manual labour, our study also aimed to assess the effort to create such a mapping: For speakers familiar with the conventions of the target representation, mapping an author-, source- or dictionary-specific orthography to an internal norm representation took between 3 and 6 working hours per source.

3.1 Normalization with FSTs

We use the Stuttgart Finite State Transducer (Schmid, 2006, SFST) library to normalize the spelling of each specific source to normalized phonological representations. In many cases, a single lemma can have multiple candidate normalizations, so that we do not operate with these directly. Instead, for any pair of lemmas that share a candidate normalization, we predict a possible link.

We base the normalized phonological representation on a specific dialect that features a small phoneme inventory (no diphthongization of long *ê*, *ô* and umlauted *ô̂*), and systematic apocope and syncope (when mapping into the norm representation, dropping a vowel can be more reliably predicted than its insertion). Several such varieties exist, both in dialects of North Saxon and East Low German, and as a point of orientation, we operate with the North Markian dialect (part of Brandenburgish in Figure 1) as documented by Pfaff (1898), Mackel (1905), Teuchert (1907) and Bretschneider (1951).[7] We first provide a mapping from source graphemes to North Markian phonemes, represented by a simplified phonological representation using ASCII characters. Where a source language grapheme has different possible readings (or different phonological mappings to North Markian), all possible interpretations are predicted. This also includes a rule to (optionally) drop *e* in all positions. In addition to context-free mappings, this also includes selected contextual assimilations and dissimilation processes, for which multi-character sequences are considered. Grapheme inventories and their mapping are established from the source dictionary in three steps:

1. Extract and split the character inventory and into vowel signs, consonant signs and other (punctuation, number) signs. This is done with a single regular expression.

2. Because many dictionaries employ digraphs or diacritics to disambiguate vowel phonemes, we extracted all continuous sequences of vowels or vowels followed by lengthening signs (*h*, *e*, *j*, *w*). For every vowel sequence, we manually verified around 50 occurrences for their pronounciation in the norm variety and provide a normalized pronounciation. We also keep note of assimilations of postvocalic vowels, and the development of vowels before *r*.

3. We then checked around 50 occurrences of every consonant as well as the apostroph (') to assess both their regular correspondence to the norm variety and the existence of possible digraphs. Special attention is paid to consonants at the end of words, and special rules for (undoing) final obstruent devoicing are added.

For a speaker familiar with SFST and the specifics of North Markian, creating an FST to normalize a dictionary in this way has been a matter of 3-6 hours, depending on the pecularities of phonology and orthography of the dialect under consideration. In addition to that, the final rulesets have been reviewed by a native speaker of North Saxon experienced in orthographical normalization across Low German dialects from different regions of Germany and the Netherlands.

3.2 Lemma Linking and Pruning

Because of ambiguities in the orthographies and complex phonological correspondences, the mapping overgenerates to some extent. As a countermeasure, we calculate confidence scores to assess the level of ambiguity in the linking. For every lemma in every dictionary, we predict all possible normalizations. For every pair of lemmas $\langle x, y \rangle$ from two dictionaries X and Y, and $x \in X$ and

[7] The approach would be applicable to *any* dialect with a similarly reduced phoneme inventory and apocope. This choice has been made because it allowed us to re-use an existing SFST implementation.

$y \in Y$, we predict a link if they overlap in one possible normalization. For the lemma y, $L_X(y)$ is the set of lemmas from X for which a link has been predicted. Link probability $P(x|y)$ is estimated as $\frac{1}{|L_X(y)|}$. We calculate confidence $c(x, y)$ of a $\langle x, y \rangle$ as the harmonic mean between the link probabilities $P(x|y)$ and $P(y|x)$: $c(x, y) = 2\frac{P(x|y)P(y|x)}{P(x|y)+P(y|x)}$.

On the basis of confidence scores, pruning is performed, in that, for every source dictionary X, only the highest-scored target links are retained. In our setting, the target dictionary is always WöWö. If there is more than one, we return the WöWö lemma with the lowest Levenshtein distance. If there is still more than one, we return the shortest WöWö lemma in order to create a bias against matches between multi-word expressions and their respective parts.[8]

3.3 Levenshtein Baseline

In the absence of training data for the respective orthographic systems under consideration here, we resort to Levenshtein distance as a baseline.[9] For performance reasons, Levenshtein distance is only calculated for the 150 most bigram-similar target lemmas for every source lemma. The linking is then pruned bidirectionally, such that for every source lemma only the most Levenshtein-similar target lemmas are preserved, and for every target lemma the most similar source lemmas.

4 Evaluation

We performed two independent evaluation experiments: evaluation against a small, manually linked glossary as a gold standard (Sect. 4.1), and evaluation of a sample of predicted links (Sect. 4.2). We employ two primary metrics, **precision** (proportion of predicted links that are confirmed), and **recall** (proportion of links in the the gold linking which are predicted). As we cannot evaluate recall from a *sample* of predicted links, Sect. 4.2 uses **coverage** in place of recall, i.e., the proportion of lemmas in the WöWö dictionary for which links are predicted (regardless if correct or not). It is to be noted that for practical application of cross-dialectal links in a lexical resource, it is essential to maintain a consistently high level of precision (at the price of recall/coverage) which, in general, should not fall below 80%. For configurations where 80% precision are achieved, we aim for maximizing coverage (resp. recall), as this leads to exhaustive linking.

We evaluate three different linking strategies: Along with FST-based linking (**FST**) and Levenshtein (**LEV**), we also compare an identity baseline (**IDENT**) which just creates a link between lemmas with identical forms. Whereas the second evaluation ran against a representative sample of Low German varieties spoken in Germany, the first evaluation (Sect. 4.1) was conducted against two varieties closer related to each other than to the North Markian dialect chosen as the basis for the normalization. In order to compensate the expected overperformance of IDENT baseline in comparison to other, less closely related dialects considered in Sect. 4.2, Sect. 4.1 employs a fourth linking strategy, **IDENT+FST** (FST-based linking with IDENT prior) which links all lemmas with identical forms, and applies FST to the remainder. As formulated above, neither FST nor LEV necessarily lead to unambiguous results. For evaluation, we restrict their predicted links to the lexicographically first candidate lemma from WöWö.

4.1 Evaluation against Wisser

The Wisser glossary is based on word list and linguistic description in Wilhelm Wisser's (1927) fairy tale collection from eastern Holstein. The glossary was manually linked with the WöWö by one of the authors, resulting in 1787 linked pairs out of 1920 overall Wisser glossary entries. Table 1 summarizes the evaluation results for the Wisser glossary. Despite WöWö (Dithmarschen / western Holstein) and Wisser (eastern Holstein) representing closely related varieties of North Saxon (which may be reflected in the high values for precision), identity matches yield relatively poor recall. This is largely due to different sets of diacritics employed to disambiguate phonemes.

[8] We enforce 1:1 correspondences between different dialects of Low German for technical reasons, as this eliminates the bias that $n : m$ correspondences could introduce into the evaluation. For practical applications of the procedure, multiple mappings of the same lexeme can be retained, if they achieve identical scores.

[9] We would expect substantially better results with weighted Levenshtein, but we have no empirical basis to set the weights accordingly. In general, a good rule of thumb for *Germanic languages* would be to penalize deviations in consonantism more than deviations in vowels, but this is a language-specific heuristic, and we would like to assess the viability of the procedure in a fully language-agnostic way. Furthermore, we also need to acknowledge that a 'naive' separation into consonants and vowels may lead to incorrect conclusions. This is partially due to diachronic phonology, assimilation and dissimilation processes, but also orthography: German-based orthographies use *h* to indicate vowel length, the Plautdietsch orthography uses *j* to mark palatalization, Dutch-based orthographies may use *j* for long *i*, etc. In none of these cases, the character is used to represent a consonant.

The Levenshtein distance (with different similarity thresholds) is more robust against such variation and outperforms the identity baseline in recall. However, its precision suffers from the failure to capture linguistically plausible replacements. As such, Wisser *ansḷan* has equal Levenshtein distance with WöWö *ansēhn* 'look at' and *ansloon* 'put up, fasten', but whereas the latter is, indeed, the correct link, the similarity with the former is chance resemblance.

In the FST-based normalization, for every Wisser word, we link the highest-scored (least ambiguous) WöWö word it shares a candidate normalization with. FSTs do not distinguish between more or less plausible (frequent) patterns and the normalization contains considerable ambiguities. As a result, FST-based linking (with different confidence thresholds) can outperform LEV in precision (not in recall), but, at least for so closely related varieties as considered here, not the IDENT-baseline. Nevertheless, FST-based linking exceeds the IDENT baseline in f-score. In terms of f-score, the overall best performance is achieved with IDENT+FST and a confidence threshold of .6. I.e., for words not linked by IDENT, we limit FST predictions to cases in which one linking direction was unambiguous, and the other had no more than two alternatives. We conclude that FST-based normalization has the potential to outperform a naive Levenshtein distance for linking, but that, for closely related dialects, a combination with plain identity matches may be a strategy to further improve the results to the necessary level of quality.

	matches (true positives)	prec	rec	f
IDENT				
(= LEV 0)	193	.878	.119	.210
LEV				
≤ 1	458	.591	.268	.369
≤ 2	612	.455	.342	.390
≤ 3	643	.410	.358	.382
≤ 4	648	.395	.361	.377
unrestricted	652	.377	.362	.369
FST				
= 1.0	206	.687	.127	.214
$\geq .6$	265	.564	.164	.254
$\geq .5$	288	.492	.180	.264
$\geq .4$	304	.451	.190	.267
unrestricted	331	.294	.206	.242
IDENT+FST				
= 1.0	445	.816	.319	.459
$\geq .6$	492	**.702**	**.355**	**.472**
$\geq .5$	512	.627	.371	.466
$\geq .4$	524	.581	.38	.459
unrestricted	546	.402	.396	.399

Table 1: Evaluation against Wisser with different thresholds.

4.2 Evaluating Predicted Links

For the evaluation of FST and LEV linking of the external dictionaries against WöWö, we randomly sampled 50 WöWö lemmas and evaluated their predicted links (using random disambiguation for m:1 links) for every dictionary. For any dictionary where not all of the previously sampled WöWö lemmas were linked, we extended the initial list of WöWö lemmas accordingly such that at least 50 predicted WöWö links were evaluated for both methods and every external dictionary. Each candidate was then manually classified into one of three categories: **exact match**, **approx-match** (partial agreement, e.g., for linking one word with a multiword expression, or with a related word, say, a morphologically derived form or a compound), and **mismatch** (including unrelated homophones).

We calculate precision in two ways, for exact-match precision, only exact matches are considered true positives, for approx-match precision, both exact and approximate matches are considered true positives. It is to be noted that the varying structures of the dictionaries linked to WöWö influenced the evaluation results: In Sass, Plattmakers, and Platt-WB, the matching rates are considerably higher because multiple inflectional word forms are grouped under the same lemma ID, and thus to be considered exact matches. This is not the case for Reuter and WWB, where, for instance, nouns and adjectives – such as *Trūe* 'loyalty' vs. *trūe* 'loyal' – are indexed separately.

The results are summarized in Fig. 6 comparing the overall match rates for LEV and FST. As the plots show, average precision in the FST-based linking consistently outperformed the baseline across all datasets. The LEV baseline achieved moderate precision but struggled especially with apocope-affected forms. For a cross-dialectal lexical resource that may also be used for consultation by humans, it is essential to guarantee a certain level of quality, for which we posit a threshold of 80% precision. In terms of exact-match precision, LEV fails to meet this threshold, but it is achieved with all FSTs with confidence ≥ 0.6. In terms of approx-match precision, this is achieved by all FSTs with confidence ≥ 0.2 as well as by LEV 0. Table 2 evaluates the coverage for these high-precision configurations, and also includes LEV 1 for comparison. Configurations with exact-match precision

	FST 1		FST ≥ 0.6		FST ≥ 0.5		FST ≥ 0.4		FST ≥ 0.3		FST ≥ 0.2		LEV 0 (IDENT)		LEV ≤ 1	
	prec	cov	prec	cov	prec	cov	prec	cov	prec	cov	prec	cov	prec	cov	prec	cov
Plattmakers	90.48%	4.66%	80.65%	**5.66%**	73.53%	6.06%	63.41%	6.36%	62.79%	**6.51%**	59.57%	6.79%	80.77%	4.88%	54.90%	9.18%
	90.48%	1243	90.32%	**1511**	85.29%	1619	82.93%	1698	81.40%	**1738**	78.72%	1812	92.31%	1304	62.75%	2450
PlattWB	85.71%	11.79%	81.08%	**15.36%**	77.78%	17.09%	70.37%	17.96%	70.18%	18.46%	66.15%	19.38%	63.64%	11.91%	59.70%	**23.61%**
	92.86%	3147	89.19%	**4100**	91.11%	4562	87.04%	4796	87.72%	4930	81.54%	5175	84.09%	3180	83.58%	**6303**
Reuter	96.88%	**5.77%**	78.72%	7.52%	69.64%	8.28%	65.63%	8.64%	65.67%	8.88%	65.22%	**9.25%**	61.76%	5.13%	55.36%	11.57%
	100.00%	**1541**	89.36%	2008	82.14%	2210	82.81%	2308	82.09%	2371	82.61%	**2470**	85.29%	1370	75.00%	3089
Sass	88.89%	15.52%	85.42%	19.11%	84.75%	20.40%	82.81%	21.12%	80.60%	**21.52%**	78.57%	22.18%	87.50%	13.58%	71.83%	**23.68%**
	91.67%	4144	91.67%	5103	91.53%	5446	92.19%	5640	91.04%	**5747**	90.00%	5923	97.92%	3625	83.10%	**6323**
WWB	96.15%	9.37%	87.18%	**12.21%**	76.00%	**13.52%**	72.73%	14.25%	66.67%	14.70%	61.11%	15.34%	85.71%	2.29%	48.94%	10.27%
	96.15%	2501	87.18%	**3260**	80.00%	**3609**	78.18%	3806	71.43%	3925	68.06%	4095	92.86%	611	68.09%	2743
average	91.62%	9.42%	82.61%	**11.97%**	76.34%	13.07%	70.99%	13.67%	69.18%	14.02%	66.12%	**14.59%**	75.88%	7.56%	58.15%	15.66%
	94.23%		89.54%		86.01%		84.63%		82.74%		80.19%		90.49%		74.50%	

Table 2: Evaluation of WöWö coverage for high-precision configurations (upper row: exact-match precision, lower row: approx-match precision; gray: $\geq 80\%$ exact-match precision, light gray: $\geq 80\%$ partial-match precision, bold: highest-coverage configuration per subset).

$\geq 80\%$ are colored gray and delineated with a solid line, configurations with partial-match precision $\geq 80\%$ are light gray with a dotted line. Within each group, the configuration with the highest coverage is marked in bold. Naturally, this coincides with lower degrees of precision. On average over all dictionaries, the best coverage for exact-match precision $\geq 80\%$ is achieved with FST ≥ 0.6, for partial-match precision $\geq 80\%$ with FST ≥ 0.2. For Sass and PlattWB, the highest coverage for partial-match precision is obtained with LEV 1, and, again, this may reflect the fact that both are dictionaries whose reference variety belongs to North Saxon and is therefore more closely related to WöWö than those of WWB or Reuter. It is to be noted that IDENT did not outperform FST or LEV as in the previous evaluation, and this may reflect the greater linguistic diversity represented by these dictionaries.

5 Results and Perspectives

We described and evaluated a methodology for creating a cross-dialectal lexical knowledge graph that consists of a core dictionary (WöWö) and its links to external, orthographically heterogeneous dictionaries of different dialects of Low German which are accessible over the web. Conceptually similar applications of these technologies to related sets of languages, dialects or dictionaries (for the same language) include, for example, the OntoLex/RDF modelling of bi-dictionaries for (primarily) the languages of the Iberian peninsula (Forcada, 2021; Gracia et al., 2018), the Bavarian dialects in Austria (Declerck et al., 2016), and a vast collection of dictionaries and related resources for Latin (Passarotti et al., 2020). However, the linking provided by resources is based on previously available conceptual or lexical knowledge, which for the case of Low German, does not exist in any electronic resource we are aware of.

Instead, our linking is based on a linguistically informed mapping of forms, without considering additional information provided by the external dictionaries. We found that the FST-based approach generally yields higher coverage than Levenshtein. This is an unsurprising result, as it effectively means that an approach with (moderate) manual effort outperforms an unsupervised approach. Yet, it is noteable that the actual manual effort is indeed low, for a speaker familiar with the language, setting up a new FST, resp., the underlying mapping tables requires approximately 3-6 hours per source dictionary.

The total number of links predicted for individual dictionaries is summarized in Tab. 3 (FST linking and unrestricted confidence scores). The resulting cross-dialectal knowledge graph is provided as a RDF/Turtle dump, and consists of multiple files (RDF Graphs) representing either a dictionary (WöWö), or its linking with an external dictionary (all other dictionaries), using URIs resolving to the original pages. The knowledge graph does not contain actual information from these dictionaries, but only lemma forms, the original URL and the confidence score. We also provide an HTML view, generated from the results of a SPARQL query. Note that this uses the URLs of the lexical entries (i.e., for external dictionaries, their native URL) as the basis for hyperlinks, so that all links can be interactively explored. For a human, this HTML file (resp., for a machine, the underlying RDF data) is capable of serving as a 'digital Rosetta Stone', linking dictionaries and mapping corresponding words

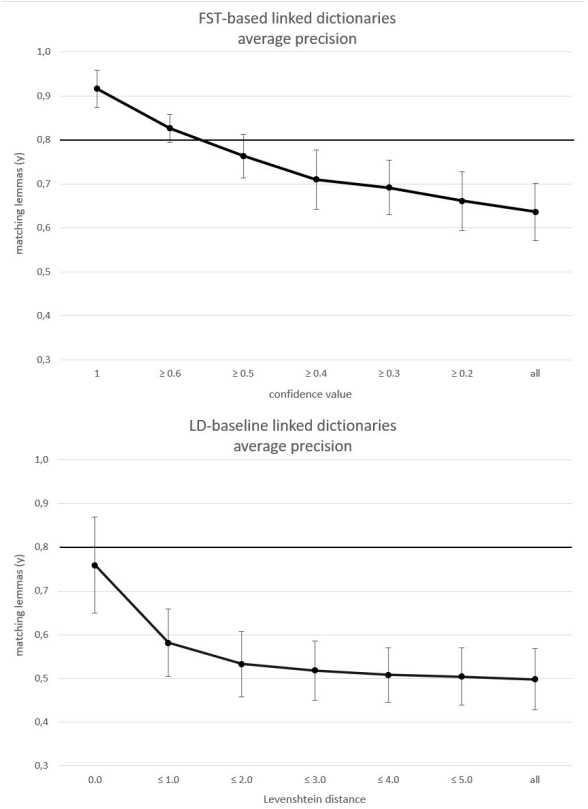

Figure 6: Average exact-match precision per source (Plattmakers, PlattWB, Reuter, Sass, WWB), computed over ≥50 WöWö links as a function of the confidence threshold for FST (top) and the distance threshold for LEV (bottom).

	linked dictionary entries		linked WöWö entries	
Plattmakers	36.30%	(2432/6700)	7.15%	(1908/26702)
PlattWB	27.81%	(6808/24479)	19.87%	(5307/26702)
Reuter	27.60%	(2834/10268)	9.60%	(2563/26702)
Sass	20.56%	(6826/33199)	22.62%	(6041/26702)
WWB	8.59%	(5841/68024)	15.76%	(4209/26702)
total			38.56%	(10295/26702)

Table 3: Resulting cross-dialectal links.

across dialects – without resorting to a standard variety or spelling.

The linking covers more than 10.000 WöWö entries. This number may appear small in comparison to the 26,702 of the WöWö in total, but to a large extent, this is due to compounds and derived forms that were included in WöWö, but not (or, at least, not as independent lemmas) in the other dictionaries because these are part of productive morphology. As such, we have 41 WöWö lemmas for *trecken* 'to pull' and its derived forms in WöWö, but only 18 of these have been linked. Further, WöWö contains a considerable number of phrasal expressions, which are not necessarily included in the other dictionaries, because these take a primary stance on documenting lexical semantics, not idiomatic expressions.

The linking method employed here primarily serves to establish a baseline for future research, our cross-dialectal dictionary serves as a testbed for a number of community standards for machine-readable dictionaries on the web in general, and for non-standardized, low-resource languages in particular. Future directions include the development of more refined methods for lexical alignment, where weights are trained on the basis of confirmed lexical link, as well as the improvement of the resources and to faciliate the further uptake of the methodologies for linking and data modelling by intensifying ties and establishing collaborations with providers of dictionaries in the realm of Low German as well as related languages and language varieties. Aside from contributing to the emerging, global network of web-accessible and interlinked lexical resources in OntoLex (Declerck et al., 2015), this also offers a possibility to improve the linking over a purely formal approach, as it then allows us to also integrate information from glosses, etc., directly, so that more traditional methods for translation inference across dictionaries (Mausam et al., 2009; Goel et al., 2022; Quadrado et al., 2023) can be used – which actually take translations and dictionary glosses into account.

Overall, we succeeded in creating an initial version of a 'Rosetta stone' for major dialects of Low German in Germany in the sense that there now is a human- and machine-readable lexical knowledge graph of (North Saxon) lemmas and their interdialectal links into other, externally hosted dictionaries. Directions for future research include using the generated links as an empirical basis for weighted Levenshtein, probabilistic FSTs and neural transliteration. Previously, these methods have not been applicable to Low German because of a general lack of parallel, cross-dialectal data.

Limitations

The WöWö dictionary is extremely rich and dense in lexicographic and linguistic information, but provided in a format tailored to human readers, which makes it virtually impossible to create an exhaustive RDF formalization. Thus, we extract and extend core aspects to include references to other dictionaries that may provide additional information, e.g., definitions. So far, we have only extracted basic information: lexical entries, written

and phonetic representations, and translations. Additional details, such as usage examples, are more challenging to extract and left for future work. Another limitation is the restriction of the resulting lexical knowledge graph to WöWö lemmas for legal reasons. Nevertheless, we consider this data a valuable contribution because it can also serve as training data for future applications of statistical or neural methods of transliteration. These can then be applied to link the respective dictionaries directly and completely, if copyright clearance can be obtained.

Improvements over our experiments would be weighted Levenshtein Distance, probabilistic FSTs or supervised neural transliteration. However, we are not aware of suitable training data. In fact, the linking we produce could represent the first instance of such data. In the absence of training data, we implement a linguistically informed normalization by means of traditional symbolic methods, and generate candidate matches between normalized source language lemmas and normalized WöWö lemmas, ranked by a confidence score that captures the level of (formal) ambiguity in $n : m$ mappings. While the phonological correspondences are well understood and uncontroversial, use of solely formal criteria is prone to link formally similar but semantically unrelated lemmas.

Because of concerns regarding legal constraints for the re-usability of external resources, the resulting knowledge graph is restricted to lemmas in the WöWö dictionary. If the linking is applied into all directions, much better coverage of the vocabulary is to be expected. We refrain from incorporating sense definitions (or glosses) of these resources into the resulting knowledge graph, as this might constitute an infringement of intellectual property, but without definitions, these links can neither be validated nor provide actual lexical information. While the resulting knowledge graph is built with LLOD technology, it does not actually constitute Linguistic Linked Open Data, as our WöWö data is linked with dictionaries whose lexical entry URIs resolve to HTML, not RDF, and most of these linked data sources are not 'open' in the sense of the Open Definition. Whenever any of these sources become accessible as Linked (Open) Data, they can, however, be seamlessly integrated.

It may seem like another limitation that we did not compare with an LLM 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. We conducted a number of such experiments with GPT-4o and focused on Reuter, the full character inventories of Reuter and WöWö dictionaries, 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 FST 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, they seem to have difficulties to generalize over the multitude of orthographies in a way that allows for transliteration. This may actually 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 literature used for training) of the model is separated from its actual transliteration capabilities (for which it lacks training data). With FSTs and an a cross-dialectal, interlinked resource, fundamental components for such a system are provided along with this paper.

Acknowledgements

We would like to thank Peter Neuber for providing us with the basis for our cross-dialectal lexical knowledge graph, as well as four anonymous reviewers for comments and feedback. The contribution of Christian Fäth has been partially supported by a travel grant by the Center for Advanced Transnational Studies (Jakob-Fugger-Zentrum) at the University of Augsburg. Janine Siewert's work was supported by the Research Council of Finland through project No. 342859 "CorCoDial – Corpus-based computational dialectology".

References

Astrid Adler and Rahel Beyer. 2017. Languages and language policies in Germany/Sprachen und Sprachenpolitik in Deutschland. In *National language institutions and national languages. Contributions to the EFNIL Conference*, pages 221–242.

Andreas Bieberstedt. 2021. Niederdeutsch als Lehrvarietät. Aspekte einer Normierung des Niederdeutschen für den Unterricht am Beispiel der Orthographie. In Birte Arendt, Robert Langhanke, and Ulrike Stern, editors, *Niederdeutschdidaktik: Ansätze–Problemfelder–Perspektiven*, pages 247–284. Peter Lang, Berlin and New York.

Christian Bizer, Tom Heath, and Tim Berners-Lee. 2009. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22.

Verena Blaschke, Hinrich Schütze, and Barbara Plank. 2023. A Survey of Corpora for Germanic Low-Resource Languages and Dialect. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 392–414.

Anneliese Bretschneider. 1951. Volkssprache der Prignitz. *Jahrbuch des Vereins für niederdeutsche Sprachforschung*, 74, p.82-98, and 75, p.62-109.

Elke Brückmann. 2025. Das Online-Wörterbuch für ostfriesisches Plattdeutsch. https://www.platt-wb.de. Accessed Jan 6th, 2025; based on Vries, Gernot de, Ostfriesisches Wörterbuch Hochdeutsch / Plattdeutsch. Oostfreesk Woordenbook Hoogdüütsk / Plattdüütsk, Leer 2000.

Marcus Buck. 2007-2024. Plattmakers Das Plattdeutsche Wörterbuch. https://plattmakers.de. Accessed Jan 15th, 2024.

Christian Chiarcos, Tabea Gröger, and Christian Fäth. 2025. Putting Low German on the map (of Linguistic Linked Open Data). In *Proceedings of the 5th Conference on Language, Data and Knowledge (LDK-2025)*, Naples, Italy.

Philipp Cimiano, Christian Chiarcos, John P McCrae, and Jorge Gracia. 2020. *Linguistic Linked Data: Representation, Generation and Applications*. Springer Nature.

Thierry Declerck, Amelie Dorn, and Eveline Wandl-Vogt. 2016. Adding Polarity Information to Entries of the Database of Bavarian Dialects in Austria. In *Proceedings of the 17th EURALEX International Congress, Ivane Javakhishvili Tbilisi University Press, Tbilisi, Georgia*, pages 654–659.

Thierry Declerck, Eveline Wandl-Vogt, and Karlheinz Mörth. 2015. Towards a pan European lexicography by means of linked (open) data. *Proceedings of eLex*, pages 342–355.

Klaas-Hinrich Ehlers. 2021. Welches Niederdeutsch unterrichten? Ein kritischer Problemaufriss vor dem Hintergrund der jüngeren Entwicklung des Niederdeutschen in Mecklenburg-Vorpommern. In Birte Arendt, Robert Langhanke, and Ulrike Stern, editors, *Niederdeutschdidaktik: Ansätze–Problemfelder–Perspektiven*, pages 29–60. Peter Lang, Berlin and New York.

Mikel L Forcada. 2021. Free/open-source machine translation for the low-resource languages of spain. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Shashwat Goel, Jorge Gracia, and Mikel L Forcada. 2022. Bilingual dictionary generation and enrichment via graph exploration. *Semantic Web*, 13(6):1103–1132.

Jorge Gracia and Daniel Vila-Suero. 2015. Guidelines for linguistic linked data generation: Bilingual dictionaries. Technical report, W3C Community Group Best Practices for Multilingual Linked Open Data (BMLOD). Final Community Group Report 29 September 2015.

Jorge Gracia, Marta Villegas, Asuncion Gomez-Perez, and Nuria Bel. 2018. The apertium bilingual dictionaries on the web of data. *Semantic Web*, 9(2):231–240.

Peter Hansen. 2025. Das Fritz-Reuter-Wörterbuch. https://www.niederdeutsche-literatur.de/dwn/index-frw.php. Accessed Jan 6th, 2025.

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

Thomas Margoni. 2016. The Harmonisation of EU Copyright Law: The Originality Standard. In Mark Perry, editor, *Global Governance of Intellectual Property in the 21st Century: Reflecting Policy Through Change*, pages 85–105. Springer International Publishing, Cham.

Peter Martens. 1979. Zum normativen Zwang der Standardsprache. Anpassung von mundartlichen Ausspracheformen und Schreibweisen an die hochdeutschen Standardsysteme. *Zeitschrift für Dialektologie und Linguistik*, pages 7–25.

Peter Martens. 2002. Zur Schreibung des Niederdeutschen. Eine Kritik der "Bremer Schreibung" in der" Niederdeutschen Grammatik" von 1998. *Zeitschrift für Dialektologie und Linguistik*, 69(H. 2):146–163.

Mausam, Stephen Soderland, Oren Etzioni, Daniel Weld, Michael Skinner, and Jeff Bilmes. 2009. Compiling a Massive, Multilingual Dictionary via Probabilistic Inference. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 262–270, Suntec, Singapore. Association for Computational Linguistics.

Carl Friedrich Müller. 1904. *Reuter-Lexikon: der plattdeutsche Sprachschatz in Fritz Reuters Schriften*, volume 19. Hesse & Becker.

Peter Neuber. 2001. *Wöhrner Wöör: Niederdeutsches Wörterbuch aus Dithmarschen ; hochdeutsch - plattdeutsch*. P. Neuber, Wöhrden.

Hermann Niebaum, Hans Taubken, Paul Teepe, Felix Wortmann, and Robert Damme. 1969-2021. *Westfälisches Wörterbuch*. Kommission für Mundart- und Namenforschung des Landschaftsverbandes Westfalen-Lippe. Wachholtz Verlag Murmann Publishers, Kiel/Hamburg. 5 Vols; digitally edited by the Wörterbuchnetz of the Trier Center for Digital Humanities, Version 01/23, https://www.woerterbuchnetz.de/WWB, accessed 2025-03-11.

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through lemmas. the lexical collection of the lila knowledge base of linguistic resources for latin. *Studi e Saggi Linguistici*, 58(1):177–212.

Hermann Pfaff. 1898. *Die Vocale des mittelpommerschen Dialects*, volume 239. A. Straube.

João Pedro Quadrado, Carlos Roberto Valêncio, Adriane Orenha Ottaiano, Geraldo Francisco D Zafalon, Luís Marcello Moraes Silva, and Angelo Cesar Colombini. 2023. Generation of new bilingual dictionaries through inference techniques supported by a graph oriented database management system. In *2023 IEEE International Conference on Knowledge Graph (ICKG)*, pages 19–26. IEEE.

Gertrud Reershemius. 2010. Niederdeutsch im Internet. Möglichkeiten und Grenzen computervermittelter Kommunikation für den Spracherhalt. *Zeitschrift für Dialektologie und Linguistik*, 77(2):183–206.

Leoš Šatava. 2019. New speakers in the context of the minority languages in Europe and the revitalization efforts. *Treatises and Documents, Journal of Ethnic Studies*, 82:131–51.

Helmut Schmid. 2006. A Programming Language for Finite State Transducers. In *Finite-State Methods and Natural Language Processing: 5th International Workshop, FSMNLP 2005, Helsinki, Finland, September 1-2, 2005, Revised Papers*, volume 4002, page 308. Springer Science & Business Media.

Andrew Charles Schwenk. 2017. *Promoting and monitoring Low German: education policies and ideologies of language in the northern German Bundesländer*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Eduard Sievers. 1875. *Der Heliand und die angelsächsische Genesis*. Lippert.

Claudia Soria, Irene Russo, Valeria Quochi, Davyth Hicks, Antton Gurrutxaga, Anneli Sarhimaa, and Matti Tuomisto. 2016. Fostering digital representation of EU regional and minority languages: The digital language diversity project. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3256–3260.

Hermann Teuchert. 1907. Die Mundart von Warthe (Uckermark). *Niederdeutsches Jahrbuch*, 33:27.

Hermann Teuchert. 1914. Zur plattdeutschen Rechtschreibung. *Zeitschrift für Deutsche Mundarten*, pages 228–237.

Heinrich Thies. 2025. SASS Plattdeutsches Netzwörterbuch. https://netz.sass-platt.de. Accessed Jan 6th, 2025.

Kathrin Weber and Timo Schürmann. 2018. Verschriftung und Normierung–niederdeutsche WhatsApp-Kommunikation innerhalb einer geschlossenen SchreiberInnengruppe. *Networx; 82*.

Wilhelm Wisser. 1927. *Plattdeutsche Volksmärchen*. Eugen Diederichs Verlag, Jena.

A Appendix

Dubenslağ	Taubenschlag			Duwenslag [1.0]	Duvenslag [1.0]	Dūwen-slag [1.0]
Dwang	Zwang	Dwang [1.0]	Dwang [1.0]	Dwang [1.0]	Dwang [1.0]	Dwang [1.0]
Dwârgenrott	Zwergenschar				Dwargenrott [1.0]	
Dwârğ	Zwerg	Dwarg [1.0]			Dwarg [1.0]	
Dwēērbâlken	Querbalken1				Dweerbalken [1.0]	
Dwēērsack	Quersack /Schultersack	Dweersack [1.0]				Dweᵃrs-sak [1.0]
Dwēērstock	Fenstersprosse					Dweᵃrs-stāke [0.67] Dweᵃrs-stok [0.67]
Dwēērweğ	Querweg	Dweerweg [1.0]				
Dwēēr|Quēēr	Quer durch den Garten					Kweᵃre [0.67]
Dârm	Darm		Darm [1.0]	Darm [1.0]	Darm [1.0]	Darm [1.0]
Dädersch|e	Täterin				Dädersche [1.0]	
Dään	Däne				Dään [0.5]	Diⁿne [0.67]

Figure 7: Interdialectal link index, HTML export (excerpt), columns from left to right showing WöWö, German translation (WöWö), Plattmakers, PlattWb, Reuter, Sass and WWB.