Putting Low German on the Map (of Linguistic Linked Open Data)

Christian Chiarcos and Tabea Gröger and Christian Fäth
Applied Computational Linguistics (ACoLi)
University of Augsburg, Germany
{christian.chiarcos|tabea.groeger|christian.faeth}@uni-a.de

Abstract

We describe the creation of a cross-dialectal lexical resource for Low German, a regional language spoken primarily in Germany and the Netherlands, based on the application of Linguistic Linked Open Data (LLOD) technologies. We argue that this approach is particularly well-suited for a language without a written standard, but with multiple, incompatible orthographies and considerable internal variation in phonology, spelling and grammar. A major hurdle in the preservation and documentation of and in the creation of educational materials such as texts and dictionaries for this variety is its internal degree of linguistic and orthographic variation, intensified by mutually exclusive influences from different national languages and their respective orthographies. We thus aim to provide a "digital Rosetta stone" to unify lexical materials from different dialects through linking dictionaries and mapping corresponding words without the need for a standard variety. This involves two components, a mapping between different orthographies and phonological systems, and a technology for linking regional dictionaries maintained by different hosts and developed by or for different communities of speakers.

1 Background

While discussing the 'digital fitness' of languages (Soria et al., 2016) with respect to their usage, dissemination and accessibility of web resources for speakers of that languages, emphasis is often put on speaker community size and the number (or existence) of resources and tools. However, such measures can be too narrow since tools like spell checkers, chatbots, MT technology, dictionaries, or plain texts may not be equally helpful to all speakers due to the language's *degree of internal diversity*, varying orthographies, and accepted standards. As a point in case, we describe an approach for creating both a machine-readable dictionary and interdialectal links for Low German (Low Saxon, ISO 639-2 `nds`), a European minority language with considerable phonological, morphological and orthographic diversity. Although Modern Low German has developed vibrant (regional) literature since about 1800, it lacks a written standard, corpora, machine-readable and interdialectal dictionaries, and, in particular, parallel texts and texts attested in more than one variety of Low German, limiting modern NLP applications. Likewise, off-the-shelf embeddings or LLMs are impractical due to inconsistent web training data.[1]

Without enforcing normalization and standardization, effective NLP support for Low German requires a "digital Rosetta stone" that allows us to integrate diverse language varieties uniformly. Although language normalization is possible, it has been a controversial topic (Christiansen, 1975), and – beyond the level of geographically confined regions – seems to be largely rejected by the speaker community. Instead, we focus on creating 'non-invasive' synergies between dialect-specific resources by linking regional dictionaries and providing a mapping routine capable of *spotting* formally corresponding words across dialects. In this paper, we primarily focus on methods to access such data for both humans and machines. While web-scale linking of dispersed data sources can be addressed using RDF and Linked Open Data technology (Cimiano et al., 2020, p.3-9), provid-

[1] We are aware of only one larger-scale experiment on using LLMs for Low German. According to public reports, however, this largely failed to achieve its preliminary goals after a 6 month piloting period, and was abandoned in August 2024, cf. https://www.ndr.de/kultur/norddeutsche_sprache/niederdeutsch/Pepper-Blog-34-Neue-wissenschaftliche-Wege,pepperblog180.html.

ing our data as Linguistic Linked Open Data (LLOD) involves a number of challenges in data modeling (of the dictionaries and inter-dictionary links), accessibility (i.e., readability for a human), and legal constraints (since many online dictionaries use proprietary licenses that restrict direct use, but linking is permitted).

Low German or Low Saxon (self-designation *Plattdüütsch*, *Nedersassisch* or *Nedersaksisch*) is a West Germanic language historically spoken in northern Germany, the Netherlands and the southern coast of the Baltic Sea. Closely related to Dutch, High German and Frisian, it has followed its own developmental trajectory since its first recorded texts from the 9th c. CE (Price, 2010) and is protected under the European Charter for Regional or Minority Languages (ECRML). Historically, (Middle) Low German served as a lingua franca around the Baltic Sea. However, with High German (in Germany) and Dutch (in the Netherlands) replacing it as the dominant languages of education, administration, and media since the 17th c., it is now considered threatened (vulnerable) (Moseley, 2010, p.25). While it still has millions of passive speakers, active speakers are far fewer and to a large extent elderly citizens (Adler et al., 2016), making intergenerational transmission a key challenge. This demands both educational material and digital tools, yet basic NLP tools such as spell checkers, machine translation, speech recognition, and text-to-speech systems are effectively absent. The fragmentation of modern Low German dialects – which have diverged greatly since the Middle Ages (Tab. 1) – further complicates digital communication. For example, some northern dialects lost the unvoiced vowels of Middle Low German (and thus parts of their morphological inventory), while others preserved them. Alongside this north-south division, there also exists an west-east division that reflects the expansion of Low German towards formerly Slavic territories during the Middle Ages, with Western dialects (historically) using a uniform verbal plural in *-(e)t*, and Eastern dialects (historically) using a verbal plural in *-en*. Dialects east of the Oder ceased after WWII but gave rise to emigrant varieties like Pomerano (a regionally recognized minority language in Brazil) and Plautdietsch (spoken by the Mennonite diaspora, predominantly in the Americas).

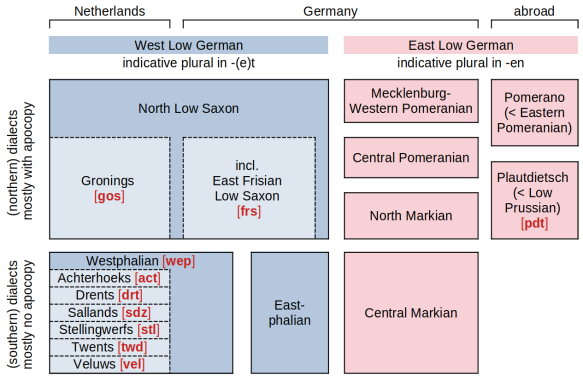

Table 1: Major dialects of Low German (ISO 639-2 `nds`), with regional ISO 639-3 codes in red square brackets.

This fragmentation makes it difficult to use the language in digital communication – reducing its visibility and usability in the modern world – and to develop tools for its Low German speakers and learners. The absence of NLP tools also hinders academic research, automated language processing, and digital content creation. Despite these challenges, Low German enjoys cultural and regional recognition. Efforts to revitalize the language include educational programs, literature, radio broadcasts, and online initiatives. These resources may play a role in transmission and revitalization of the Low German language, and indeed, this is what we see for other minority languages all over the world. However, to preserve Low German, more work is needed to integrate it into digital spaces. Developing NLP tools, expanding online resources, and boosting media presence are crucial for its survival as a living language. Currently, fundamental NLP resources are lacking, including corpora (Siewert et al., 2021), parallel corpora, and machine-readable dictionaries.

A *machine-readable dictionary (MRD)* is a lexical resource structured for computational use rather than human readability. Unlike traditional dictionaries, MRDs are formatted in a way that allows software applications to process and analyze linguistic data efficiently. They store information such as word meanings, grammatical properties, pronunciations, and translations in a structured manner to facilitate the development of downstream applications. For low-resource languages, MRDs play a crucial

role in developing foundational NLP technologies. In particular, this is the case for language varieties that have been the subject of linguistic research in the past (so that word lists or dictionaries are available), but that have been largely neglected by NLP or corpus linguistics (so that no digital corpus data is available). We are unaware of any existing comprehensive Low German MRD, aside from isolated Low German terms in foreign-language editions of DBnary (Sérasset and Tchechmedjiev, 2014) (which is crowd-sourced and inconsistent). This paper describes the development of a prototypical interdialectal MRD for Low German, consisting of two parts, a core built from a North Low Saxon dictionary of Dithmarschen (Neuber, 2001, further WöWö), republished in 2019 as *Frie' Woor* 'freeware' digital-born DOCX and PDF files. To the best of our knowledge, this is the only digital dictionary of a regional variety of Low German in Germany for which free redistribution is explicitly allowed.[2] This is complemented by interdialectal links, derived from various digital dictionaries, though all are designed for human consumption, and not for subsequent use in natural language processing. In addition, most of these are copyright-protected, either explicitly or by default copyright (if copyright is undeclared). Our approach can, however, be extended to other Low German dictionaries and dialects if copyright can be secured.

A key technology for building structured and interoperable MRDs is *OntoLex-Lemon*, an RDF vocabulary designed for representing lexical and semantic data on the web (McCrae et al., 2017). OntoLex allows lexicons to be linked to external knowledge bases and other linguistic resources, enhancing interoperability. It uses the Resource Description Framework (Beckett et al., 2014, RDF), a W3C standard to provide a flexible, graph-based data model that enables rich semantic annotations and structured linguistic relationships. Together, these technologies ensure that dictionaries for low-resource languages are not isolated but can be *integrated into broader linguistic ecosystems*, facilitating cross-linguistic research and NLP. By leveraging OntoLex and RDF, MRDs for low-resource languages can be built in a way that supports automated processing, encourages digital preservation, and enables their incorporation into modern NLP applications. These technologies make it easier to link lexical resources across languages, ensuring that low-resource languages gain better representation in computational linguistics and digital tools. As such, OntoLex has been a cornerstone for integrating lexical data into the Linguistic Linked Open Data cloud (Declerck, 2018).

The *Linguistic Linked Open Data (LLOD)* cloud (Chiarcos et al., 2011; Pareja-Lora et al., 2019; Cimiano et al., 2020) is an interlinked network of linguistic resources following Linked Data principles (Bizer et al., 2009).[3] It provides a semantic web-based infrastructure for representing and integrating linguistic data, including lexicons, corpora, terminologies, and ontologies. A key advantage of the LLOD approach is its ability to connect diverse linguistic datasets, making them accessible for computational use. The LLOD cloud benefits low-resource languages by linking their limited linguistic data to richer datasets, fostering NLP development and linguistic research. By structuring linguistic resources using open standards, the LLOD cloud contributes to the creation of multilingual and interoperable NLP systems, supporting tasks such as machine translation, semantic search, and corpus analysis. For languages with scarce and scattered data, LLOD is vital for digital preservation and computational access to linguistic knowledge.

2 Wöhrner Wöör (WöWö)

2.1 Overview and Digital Evolution

The *Wöhrner Wöör* is a Low German dictionary from the Dithmarschen region (North Low Saxon), compiled by Peter Neuber (born 1939 in Szczecin), a linguist and educator. The dictionary was created with the goal of documenting and preserving the traditional vocabulary and expressions of Plattdeutsch while simultaneously adapting the language to modern contexts. Beyond recording historical terms,

[2]There also is a multi-dialectal Low German Wiktionary under CC BY-NC-SA. However, this is crowd-sourced, and thus orthographically inconsistent and not considered here.

[3]The native home of the LLOD cloud diagram is https://linguistic-lod.org/. Since 2018, it has been formally integrated into the LOD cloud diagram and is currently provided as a separate LOD subcloud under https://lod-cloud.net/#linguistic.

Neuber sought to introduce neologisms for contemporary concepts that previously lacked Low German equivalents, integrating them into the lexicon.

First published in 2001 in Wöhrden, the *Wöhrner Wöör* consists of 699 pages and serves as a German-to-Low-German reference work specific to the Dithmarschen dialect (Fig. 1). Following its initial print release, the dictionary has undergone continuous expansion, with subsequent versions distributed exclusively in digital formats such as Microsoft Word and PDF. The latest version, titled *Ditschiplatt: Wöhrner Wöör* from January 2019 is accessible online.[4]

Despite a remarkable level of detail and complex structure, the *Wöhrner Wöör* remains primarily a resource for human readers, lacking structured machine-readable representations that would facilitate its use in NLP applications. Thus, our goal was to convert the *Wöhrner Wöör* into an RDF-based format following the OntoLex-Lemon model to ensure interoperability with other lexical datasets and enable the dictionary's inclusion in the LLOD cloud, paving the way for broader computational applications and cross-linguistic research.

2.2 Converting the WöWö

Converting the *Wöhrner Wöör* into an MRD posed a significant challenge due to its highly fragmented DOCX format. The extensive use of diverse fonts, colors, and sizes—each encoding different functions—meant that the underlying text information was split into numerous small fragments within the Office Open XML format. This complexity required a multi-stage processing pipeline via Python for extraction, merging, and transformation of the text information:

1. **Extracting relevant data from XML**
   First, the verbose XML structure of the Word document is parsed using Python's `xml.etree`. Each text run (`<w:r>`) is extracted along with its formatting metadata (font, color, and size), leveraging XML namespaces to accurately retrieve `<w:t>` (text) and `<w:rPr>` (formatting) elements. This step generates a preliminary `DataFrame` stored as a raw CSV file.

2. **Merging Consecutive Text Blocks**
   Due to fragmentation, consecutive text blocks with identical formatting are merged. A Python script iterates through the DataFrame, combining segments that share the same color and size. This merging produces a more coherent CSV that better reflects the original document's logical layout.

3. **Structuring the Data into a Lexical CSV**
   With the merged text available, the next step involves classifying and extracting entries into five columns, depending on the corresponding formatting:

   (a) **High German Main Lemma**
   (b) **High German Sublemma**
   Potential subentries per lexical entry.
   (c) **Low German Translation**
   (d) **Low German Additions**
   Additional grammatical information – mainly plural forms – that has the same formatting as the corresponding Low German lexical entry.
   (e) **Low German IPA Information**
   Low German phonetic transcriptions.

   This structured CSV serves as the foundation for converting the data into RDF.

4. **Generating RDF (Turtle Format)**
   Separate Python scripts convert the structured CSV data into RDF (Turtle):

   (a) **High German Entries:** Entries are first grouped by main lemmas. The script converts them into `ontolex:LexicalEntry` nodes, each with its own `ontolex:LexicalSense`. Additional information, such as synonymous terms or usage examples – but mostly plural information or alternative spellings (e.g., variations in single vowels) – is included as `ontolex:otherForm`. In the case of alternative spellings or plural information, these additions are usually not full words but only the modifications, such as the suffix '-s'.
   A custom property `neuber:subEntry` links to related sublemmas. For all

[4] https://ditschiplatt.de/woehrner-woeoer/

Figure 1: Excerpt of the first entries under ‘A’ from the beginning of the lexical part of the *Wöörner Wöhr* dictionary in docx format.

existing sublemmas, individual lexical entries with their own lexical senses are generated in a similar way.

(b) **Low German Translations:** The Low German translations are processed into lexical entries, each with its own lexical sense. If available, IPA notation is incorporated into the canonical form as `ontolex:phoneticRep`.

(c) **Linking Translations:** Finally, unique `vartrans:Translation` entries are generated to link source senses (High German main or sublemmas) with their corresponding target senses (Low German translations).

5. **Post-Processing**
   The generated Turtle files are further refined using a regex-based clean-up. This post-processing step removes unnecessary whitespaces, replaces dashes with underscores, and normalizes punctuation to ensure that the RDF output adheres to the required naming conventions and syntactic standards.

This comprehensive pipeline successfully transforms the fragmented DOCX format of the *Wöhrner Wöör* into a coherent RDF dataset (cf. Fig. 2), aligning the dictionary with the Ontolex-Lemon model, and thus builds a baseline for LLOD integration. So far, this extraction process has focused on retrieving the most essential information – lexical entries, written and phonetic representations, and their corresponding translations. However, the *Wöhrner Wöör* contains numerous additional details for each entry, such as references and usage examples, which are more challenging to extract due to the complexity of the fragmented format.

3 Linking the WöWö

A number of online dictionaries for Low German are available, but usually not under permissive licenses. As a result, we focus on the *WöWö* dictionary as our primary dataset, and do currently not provide Linked Data editions of other Low German dictionaries. However, these are accessible online, usually 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 create a machine-readable ‘entry point’ (i.e., an index) in RDF. As we do not use any specific information from the dictionaries other than the existence of a lemma, we assume that this information does not meet the threshold of originality legally required for copyright to apply Margoni (2016), so that these LOD indices to other Low German dictionaries can be published as addenda to the *WöWö* dataset regardless of the licensing situation of the full data sets. However, should these respective resources be ever served as Linked Data or

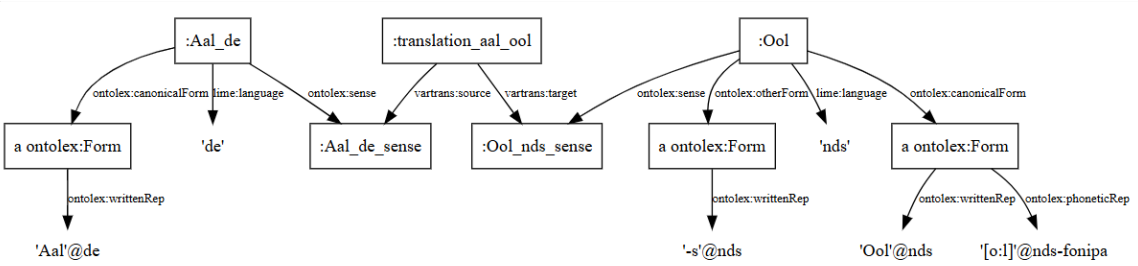

Figure 2: Resulting RDF graph for the entry *Aal* 'eal'.

be made accessible under a more permissive license, the information from the indices/links we provide can be seamlessly integrated into the respective dictionaries.

3.1 External Datasets

The dictionaries that we link with the *WöWö* are perfect silos, in the sense that they 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 existing platforms have been *designed* to provide interdialectal links, resp., links between different dictionaries, but they only provide links *within* the respective ecosystem, whereas we pursue an open, extensible approach capable of integrating *any* piece of information accessible on the web.

- The Trier Wörterbuchnetz[5] is an online platform that provides online access to dictionaries of historical and regional vernaculars, predominantly from Germany, including dictionaries for historical stages and dialects of German. Among Latin, Ladin, Uighur and Russian, it also comprises a major dictionary of the Westphalian dialect of Low German. Overall, the Wörterbuchnetz builds on mature XML technologies to provide human-readable content, and there also is an API that can be used to retrieve a lemma list (but not the content itself). Within the Wörterbuchnetz, hyperlinks are limited to resources provided by the Wörterbuchnetz itself – and at the moment, none of these are concerned with Low German, but if these should ever emerge, our linking technology may be trivially expanded to them as well as to other Wörterbuchnetz data, if a phonological mapping can be established.

- The Digitales Wörterbuch Niederdeutsch (DWN)[6] by Peter Hansen is a website that provides access to a 'basis' Low German dictionary (adopting spelling rules developed for North Low Saxon), a dictionary for Mecklenburgian-Western Pomeranian as well as custom dictionaries for selected authors (Klaus Groth, Fritz Reuter and John-Brinckman Wörterbuch). Each dictionary comes with its own search dialog, and little is known about the technical details, as only a human-readable HTML rendering is accessible. Within each dictionary, lemmata are linked across these datasets with HTML links. We presume that this uses standard SQL technology. Again, no links to external resources are being provided. As the content is copyright-protected, we decided to work only with the Reuter dictionary based on (Müller, 1904), as this goes back to a print dictionary in the public domain. We did not exploit the interdialectal links provided by the DWN, nor did we use any of its original content.

- Plattmakers[7] is an online aggregate dictionary with 22.000 entries provided in a single, searchable database, and developed by Marcus Buck. It provides its content in human-readable fashion, and individual entries are equipped with maps and links to the source literature. Plattmakers is a private website, but some details about its implementation are provided,[8]

[5] https://woerterbuchnetz.de/

[6] https://www.niederdeutsche-literatur.de/dwn/

[7] https://plattmakers.de/de

[8] https://plattmakers.de/de/faq

indicating that it is based on a relational database backend, and supported by automated normalization routines similar to those described below. Unlike DWN and Wörterbuchnetz, Plattmakers lemma URLs provide machine-readable metadata in JSON-LD, so that its content *can* be processed and evaluated in conjunction with *WöWö* information. At the same time, it is copyright-protected, so that we do not work with any Plattmakers information except for URL and lemma form.

Overall, we link five online dictionaries, covering the main branches of modern Low German, each identified with language combine ISO 639-2/-3 codes with Glottolog identifiers:[9] in the BCP47 'private use' section:

Plattmakers (for North Low Saxon/North Hanoveranian, `nds-x-nort3307`).

WWB Westfälisches Wörterbuch from Wörterbuchnetz (for Westphalian, `wep`).

Twents Twents Woordenboek by Goaitsen van der Vliet (2025), available for online search under `https://twentswoordenboek.nl` and published under CC BY-NC-SA (`twt`, a Dutch Westphalian dialect).

Reuter dictionary from DWN (for Mecklenburgian, resp., East Low German in Germany, `nds-x-meck1239`)

Plautdietsch (Mennonite Low German) dictionary by Herman Rempel and the Mennonite Literary Society (1984-1995), `mennolink.org` (1998-2006), and Eugene Reimer (2006-2007), published under CC BY-SA[10] (for emmigrant varieties of East Low German, `pdt`).

3.2 Data Retrieval and Processing

Creating an LOD index for a dictionary typically requires to retrieve a list of lemmas, e.g., by crawling its content in order to extract lemma forms and lemma URL which are then stored in a TSV file. From these initial TSV files, we then create an extended TSV file that

aarvn	https://twentswoordenboek.nl/lemmas/id/AAOF	Gröte Ârfen	http:/
_	_	Grône un Gele Ârfen	http:/
_	_	Graue Ârfen	http:/
_	_	höge Ârfen	http:/
_	_	siede Ârfen	http:/
_	_	Ârv|Ârben	http:/
_	_	ârben	http:/
aarvnsoep	https://twentswoordenboek.nl/lemmas/id/AAOG	Ârfensupp	http:/
abonneern	https://twentswoordenboek.nl/lemmas/id/AAPA	abonnëren	http:/
acht	https://twentswoordenboek.nl/lemmas/id/AAQB	Acht	http:/
_	_	(sô) hèn no (Klock) acht	http:/
_	_	in acht Dooğ	http:/

Figure 3: Linked TSV file except, Twents (left) to *WöWö* (right)

adds two additional columns, the lemma form in *WöWö* (for verification), and the *WöWö* URL (for the actual linking). All the dictionaries that *WöWö* will be linked with comprise form-level information, only, linking is grounded on *formal agreement* only, so that in most cases, there are many-to-many relationships between dictionary lemmas and *WöWö* entries (cf. Fig. 3).

This data is diverse in phonology and orthography, so that formal linking must not rely on mere identity. Instead, we use Finite State Transducers to generate hypothetical normalizations against one specific variety of Low German and then generate candidate links for lemmas from different dictionaries for which identical forms are generated. We normalize towards North Markian, an East Low German variety that resembles the North Low Saxon dialects of *WöWö* and Plattmakers in exhibiting both a reduced inventory of diphthongs and the systematic dropping of unstressed Middle Low German *e* (apocope, syncope). The mapping is implemented with the Stuttgart FST library (Schmid, 2006, SFST), using the sound correspondences established by Pfaff (1898), Teuchert (1907) and Mackel (1905). As for the effort required to implement a mapping, this normally took about a day per dataset. Low German dialects don't deviate much in their consonants, but coonlynsiderably both in their vowel inventories and the spelling of vowels. The normalization is not exposed to the user, but used internally, only: We predict a candidate link for every pair of lemmas that have at least one normalized form in common.

For the RDF export, we calculate the confidence of a link $\langle x, y \rangle$ as the harmonic mean between the linking probabilities $P(x|y)$ and $P(y|x)$, with $P(x|y)$ and $P(y|x)$ estimated from the the total of many-to-many candidate

[9] `https://glottolog.org/`

[10] `https://ereimer.net/plautdietsch/pddefns.htm`

links for the lemmas x and y, respectively. In the RDF export, we only include the most probable links.

3.3 RDF Representation

In the RDF export, we only include the most confident link, by default. For any given link $\langle x, y \rangle$, the confidence score $c(x, y)$ is calculated as $c(x, y) = 2\frac{P(x|y)P(y|x)}{P(x|y)+P(y|x)}$. If more than one match with the same score is found, we return the one with lowest Levenshtein distance. If this is not umambiguous, we return the shortest target URL in order to create a bias against partial matches. For every external dictionary, we create one lexical entry per source URL, and provide the lemma form as its canonical form. These lexical entries are then linked with *WöWö* URLs.

We produce linkings in two different flavours. The condensed format only conveys a `lexinfo:geographicalVariant` link between two lexical entries. This compact format is well-suited for downstream applications where only the link itself is processed, but it omits provenance and confidence information. Unlike the reified data described below, this is also OWL2/DL-compliant.

As there is no manual quality control involved here and the automated linking procedure creates many n:m correspondences, it is, however, preferred to provide the confidence scores, as well, for which we adopt a reified representation inspired by Gillis-Webber (2023), with a `vartrans:LexicalRelation` object that `vartrans:relates` an external lexical entry with a lexical entry from *WöWö* and that uses `lexinfo:category` to indicate the type the of relation. There are, however, no exactly corresponding concepts in lexinfo to indicate the type of relation, so that, instead of an individual, we resort to `lexinfo:geographicalVariant`, again. However, this is an object property, not an individual, the resulting data is thus propelled into the semantic space of OWL2/Full. Every reified link is complemented with a numerical confidence score. Due to the lack of a standard vocabulary for confidence scores in RDF or LexInfo, we adopt `rdf:value` for the purpose, but this is semantically underspecified.

For linking WöWö *Ool* with the Twents dictionary, we arrive at the graph in Fig. 4. The

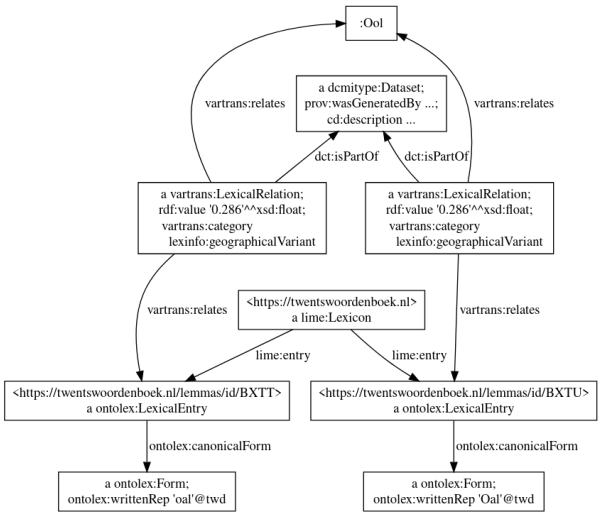

Figure 4: Reified `lexinfo:geographicalVariant` links between *WöWö Ool* 'eal' and Twents dictionary

	$c = 1.0$	$c \geq 0.65$	$c \geq 0.5$	total
Plautdietsch	834	1,260	1,416	3,665
Plattmakers	1,306	1,676	1,895	2,433
Reuter	1,571	2,107	2,375	2,835
Twents	1,641	3,200	4,775	10,149
Westphalian	2,472	3,585	4,259	5,761

Table 2: *WöWö* links with different dictionaries, filtered by confidence scores

lexical entry `:Ool` is the *WöWö* lexical entry, the individual links are formally associated with a dataset object, like the individual dictionary entries are associated with their source URL that is defined as a `lime:Lexicon`. However, as we only provide a shallow wrapper around the original source document, and because the URLs will not resolve to machine-readable information anyway, we bundle both linking information and the lexical entries drawn from `https://twentswoordenboek.nl` in a single file.

4 Querying Interdialectal Links

For evaluation, we used a single SPARQL SELECT query to retrieve all *WöWö* lemma forms, their URL, (a concatenation of) their German translations, as well as aggregates (concatenations) of lemmas, confidence scores and URLs for all external dictionaries (Appendix A). With this query, this information can be conveniently retrieved and exported to HTML. Both the query and its results are bundled with the release of our data and a snippet of the

HTML output is shown in Fig. 5. 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.

On this basis, we conducted a qualitative evaluation for 50 randomly sampled links for Reuter and WWB (Tab. 3): Overall, we found the majority of links (82% for Reuter, 64% for WWB) to represent exact or approximative matches, and in line with relative proximity of Reuter and *WöWö* varieties, with much better results for Reuter. One major factor for the high number of mismatches is that both North Low Saxon (*WöWö*) and Mecklenburgian (Reuter) drop unstressed Middle Low German e (apocope and syncope), whereas the Westphalian varieties (WWB and Twents) normally maintain it. As we cannot reliably distinguish stressed and unstressed syllables, the Westphalian (WWB and Twents) normalization allows to omit *any* e, so that words like Twents *efn* 'respectable' and *ven* 'swampy meadow' include the same (possible) normalizations and can thus be easily confused. We use Levenshtein distance as an additional disambiguating factor along with normalization-based confidence, and dialects with apocope and syncope are likely to yield forms that are more similar to WöWö, whereas the degree of variation (and the Levenshtein distance) is generally greater to dialects without apocope.

By approximative matches, we mean that either one of the words in a multi-word expression is identical, e.g., *Block Speck* 'chunk of bacon' with Plattmakers *Block* 'block, chunk, large piece', or that it involves a more or less transparent shift of meaning, e.g., *Ool* 'eal' with Twents *Oal* (derogative nickname for persons notorious for speaking glibly), based on Twents *oal* 'eal' (which is also linked). The varying structures of the dictionaries linked to *WöWö* influence the evaluation results. In Sass, Plattmakers, and Platt-WB, the matching rates are considerably higher because different word forms of the same root (nouns, verbs, adjectives, and adverbs) are grouped under the same lemma ID. This is not the case for Reuter and WWB where, for instance, nouns and adjectives–such as Trūe (noun) vs. tr e (adjective)–are indexed separately. In such cases, the same confidence score is assigned,

	Reuter	WWB
match	0.66 (33/50)	0.59 (29/50)
approx. match	0.16 (8/50)	0.06 (3/50)
mismatch	0.18 (9/50)	0.36 (18/50)

Table 3: Qualitative evaluation for 50 *WöWö* lemmas

but if the adjective is selected for the noun entry *Tru* in *WöWö* by chance, it only results in an approximate match. The category of mismatches also includes homophones, e.g., WWB *Ō¹st* 'branch' and *Ō²st* 'east', which are historically unrelated yet formally identical (in some varieties, at least) and can thus not be disambiguated by any method of form-based matching. We conclude that our formal linking method represents a reasonable baseline for future research to improve upon. In particular, such improvements can be achieved if meaning relations (i.e., the glosses, definitions and translations in the respective dictionaries) are taken into account. For the time being, we recommend downstream applications for the cross-dialectal linking to operate with high-confidence links, only, i.e., cases in which the lack of ambiguity in the formal agreement indicates a reliable link. For the cautious user, we recommend a confidence threshold of > 0.5, as this entails that at least one direction of the linking was formally unambiguous.

The total number of links predicted for individual dictionaries is summarized in Tab. 2, reporting only the most confident link for every source dictionary lemma. In total, the linking covers 8,001 *WöWö* entries, thus conforming these to be lemma forms. This number appears to be small in comparison to the 26,713 lexical entries of *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. As such, we have 41 lexical entries for *trecken* 'to pull' and its derived forms in *WöWö*, but only 18 of these have been linked. The reason is not so much that words such as *rantrecken* 'to pull here', *rintrecken* 'to pull inside', *roptrecken* 'to pull up there', *rövertrecken* 'to pull over', *rumtrecken* 'to pull over', or *ruuttrecken* 'to pull out' don't exist in the other varieties, but they haven't necessarily been included in the other dictionaries because their formation follows a regular

Dụ̀benslaǧ	Taubenschlag			Duwenslag [1.0]	doevnslag [1.0]	Dūwen-slag [1.0]
Dwang	Zwang	Dwank [1.0]	Dwang [1.0]	Dwang [1.0]		Dwang [1.0]
Dwârǧ	Zwerg	Dwoaj [1.0]	Dwarg [1.0]		dwearg [1.0] dwearg [1.0]	
Dwēērsack	Quersack /Schultersack		Dweersack [1.0]			Dweᵉrs-sak [1.0]
Dwēērstock	Fenstersprosse					Dweᵉrs-stāke [0.67] Dweᵉrs-stok [0.67]
Dwēērweǧ	Querweg		Dweerweg [1.0]			
Dwēēr|Quēēr	Quer durch den Garten				kweer [1.0]	Kweᵉre [0.67]

Figure 5: Interdialectal link index, HTML export, columns from left to right: *WöWö*, *WöWö* translation, Plautdietsch, Plattmakers, Reuter, Twents, WWB

and productive morphological pattern and they don't convey a semantic meaning that cannot be deduced from its parts. In fact, any locative adverb can be combined with *trecken* and similar verbs of motion. The same holds true for nominal compounds, which are about as productive as in High German, but are normally not included in the other dictionaries unless they have special semantics that cannot be derived from its parts.

5 Discussion and Outlook

We propose a method for creating a cross-dialectal lexical resource for Low German using LLOD technologies. This approach is particularly suited to a language that lacks a standardized written form, exhibits multiple conflicting orthographies, and shows significant internal variation in phonology, spelling, and grammar. We provide a conversion of the *WöWö* dictionary of the Dithmarschen dialect of North Low Saxon into RDF and use this as a lexical backbone. In a second processing step, this was enriched with cross-dialectal links based on formal agreement of *WöWö* lemmas with lexical entries from dictionaries of 5 other Low German dialects. This data is provided as RDF data, with three files representing the original *WöWö* and one RDF file per external dictionaries. These RDF files define lexical entries and their respective canonical forms, but they do not provide additional details beyond the location of the corresponding lexical entry on the web – the URI of the lexical entry is the URL of the underlying lemma. With the external dictionaries not providing an RDF view on their content, this is not actually linked data, as these URIs do not resolve to machine-readable data, but it is possible to query the graph and to provide a tabular export that not only includes (excerpts of) *WöWö* information, but also links with external dictionaries.

We provide an HTML view on this tabular export, and for a human, this HTML file (resp., for a machine, the underlying RDF data) is actually capable of serving as a "digital Rosetta Stone", linking dictionaries and mapping corresponding words across dialects – without resorting to a standard variety or spelling (which, for the case of Low German, does not exist). Aside from supporting speakers and learners in their exploration of interdialectal differences and similarities, this approach also enables new applications in the technical realm: Since there are no cross-dialectal parallel texts for Low German, linking dictionaries could facilitate the induction of multidialectal word embeddings – and, building upon that, multidialectal contextualized embeddings. Each of the dialects examined here has its own literary tradition, written in different orthographies.

While our linking method primarily serves to establish a baseline for future research, our cross-dialectal dictionary provides 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. We observed a number of potential gaps in the existing OntoLex vocabularies.

1. As our interdialectal links are created by heuristic means, we would like to be able to express to what extend a user can rely on the information conveyed by a link. This includes *candidate links* (with a property such as '`...:possibleMatch`'), but also the possibility to mark a link as an (un)verified hypothesis.

2. It would be good to have a standard vocabulary for confidence in OntoLex, resp.,

LexInfo. PROV-O (Jing, 2015) does not provide a codified vocabulary for expression confidence scores, in fact, the PROV-O documentation has an example that uses a *local* property to provide that information, and PROV-O users have resorted to their own properties, too, e.g., `nif:taIdentConf`, `nif:taClassConf`, or `nif:confidence` in the NLP Interchange Format.[11] But these properties are designed for a different purpose (linguistic annotation) and should not be applied to lexical linking. It should be noted that confidence scores are a recurring component of lexical resources, but apparently, no standard practice has been established in that regard. More generally, this is an intensely researched problem in the RDF world, and one of the key motivations behind RDF-star (Rupp et al., 2024).[12]

3. Lexinfo currently does not support the reification of `lexinfo:geographicalVariant` (and its sibling properties). As we have to point with `lexinfo:category` to an object property, we move the entire dataset out of the realm of OWL2/DL and into OWL2/Full. As a result, standard reasoning techniques cannot be applied to the resulting lexical knowledge graph. It would be ideal, if there would be an individual with a similar meaning.

In addition to this, we found some solutions for apparent OntoLex gaps, and these may even entail future simplifications: As such, there is an apparent gap of a counterpart of translation sets for relations other than translations in OntoLex-VarTrans, but we found an acceptable work-around in `dct:Dataset`, and we would suggest this as a best practice for other types of lexical-semantic relations, as well. Yet, to align this approach better with the current treatment of translation(set)s, one may consider to re-define `vartrans:TranslationSet` as a subclass of `dct:Dataset` (and `vartrans:trans` as a subproperty of `dct:hasPart`) and to motivate it as such in a future revision of the VarTrans module. This would be a backward-compatible revision that comes without any additional overhead (i.e. newly introduced concepts). A more radical alternative would be to deprecate `vartrans:TranslationSet` and to refer `dct:Dataset`, instead.

Overall, we succeeded in creating our 'Rosetta stone' for representative varieties of Low German in the sense that there now is a human- and machine-readable lexical knowledge graph of (North Low Saxon) lemmas and their interdialectal links into other, externally hosted dictionaries. However, while we were using standard LLOD technologies to implement this interdialectal linking, we did not actually provide Linguistic Linked Open Data. Our *WöWö* data uses resolvable URIs, but it is linked with dictionaries in HTML, but not RDF. Further, most of these linked data sources are not actually 'open' in the sense of the Open Definition. But our work represents a first step towards putting Low German on the map of Linguistic Linked Open Data, and a proof-of-principle of its capabilities. A future direction may thus be to encourage or to support the colleagues developing Wörterbuchnetz, DWN, and other platforms, to embrace RDF technologies, and then, to really create an interdialectal, distributed meta-dictionary of Low German, and to facilitate the development of technologies and resources that benefit *all* its varieties in their entirety.

The RDF data is publicly available from the NDS Spraakverarbeiden organization at GitHub and from `https://nds-spraakverarbeiden.github.io/linked-nds-dictionaries/`.

Note that after conversion, we had to drop the Twents lemma URLs from the HTML release, because we found these to be unstable. (The data is still included in RDF, and can be re-built from the repository any time.) We actually see this as a call to arms for the promotion of Linguistic Linked Data and Open Data, as here, developer convenience and copyright restrictions force us to exclude a potentially important linguistic data set (and a speaker community) from interdialectal lexical resources and technical solutions developed on this basis.

[11] `https://nif.readthedocs.io/en/latest/prov-and-conf.html`

[12] `https://www.w3.org/groups/wg/rdf-star/`

References

Astrid Adler, Christiane Ehlers, Reinhard Goltz, Andrea Kleene, and Albrecht Plewnia. 2016. *Status und Gebrauch des Niederdeutschen 2016*. Institut für Deutsche Sprache, Mannheim.

David Beckett, Tim Berners-Lee, Eric Prud'hommeaux, and Gavin Carothers. 2014. RDF 1.1 Turtle. Technical report, World Wide Web Consortium.

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.

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. 2011. Towards a Linguistic Linked Open Data cloud: The Open Linguistics working group. *Traitement automatique des langues*, 52(3):245–275.

Heinz C. Christiansen. 1975. *Reuter und das Plattdeutsche*, pages 15–30. J.B. Metzler, Stuttgart.

Philipp Cimiano, Christian Chiarcos, John P McCrae, and Jorge Gracia. 2020. *Linguistic Linked Data. Representation, generation and applications*. Springer, Cham, Switzerland.

Thierry Declerck. 2018. Towards a Linked Lexical Data cloud based on OntoLex-Lemon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan*, pages 7–12.

Frances Gillis-Webber. 2023. Refinement of the classification of translations. Extension of the vartrans module in OntoLex-Lemon. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 37–48, Vienna, Austria.

Ni Jing. 2015. A PROV-O based approach to web content provenance. In *2015 International Conference on Logistics, Informatics and Service Sciences (LISS)*, pages 1–6. IEEE.

Emil Mackel. 1905-1907. Die Mundart der Prignitz. *Niederdeutsches Jahrbuch*, 31-33.

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.

John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The Ontolex-Lemon model: Development and applications. In *Proceedings of the Fifth Biennial Conference on Electronic Lexicography (eLex 2017)*, pages 19–21, Leiden, Netherlands.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. UNESCO Publishing, Paris. 3rd edn.

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

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

Antonio Pareja-Lora, Barbara Lust, Maria Blume, and Christian Chiarcos. 2019. *Development of linguistic linked open data resources for collaborative data-intensive research in the language sciences*. The MIT Press.

Hermann Pfaff. 1898. *Die Vocale des mittelpommerschen Dialects. Inaugural-Dissertation zur Erlangung der philosophischen Doctorwürde der Universität Leipzig*. A. Straube, Labes.

Timothy Blaine Price. 2010. *The Old Saxon Leipzig Heliand manuscript fragment (MS L): New evidence concerning Luther, the poet, and Ottonian heritage*. Ph.D. thesis, University of California, Berkeley.

Florian Rupp, Benjamin Schnabel, and Kai Eckert. 2024. Implementing data workflows and data model extensions with RDF-star. *The Electronic Library*, 42(3):393–412.

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*, page 308.

Gilles Sérasset and Andon Tchechmedjiev. 2014. DBnary: Wiktionary as linked data for 12 language editions with enhanced translation relations. In *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, pages 67–71, Reykjavik, Iceland.

Janine Siewert, Yves Scherrer, and Jörg Tiedemann. 2021. Towards a balanced annotated Low Saxon dataset for diachronic investigation of dialectal variation. In *17th Conference on Natural Language Processing (KONVENS 2021)*, pages 242–246, Düsseldorf, Germany.

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, Portoroz, Slovenia.

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

A Sample Query

The following SPARQL query was used to construct an integrated HTML view over WöWö lexical entries, their translations and and their respective links.

```
PREFIX vartrans: <http://www.w3.org/ns/lemon/vartrans#>
PREFIX ontolex: <http://www.w3.org/ns/lemon/ontolex#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?woewoe ?gloss ?pdt ?plattmakers ?reuter ?twents ?wwb
WHERE {
 ?le_woewoe a ontolex:LexicalEntry.
 FILTER(contains(str(?le_woewoe),"/woewoe/"))
 ?le_woewoe ontolex:canonicalForm/ontolex:writtenRep ?woewoe_form.
 BIND(concat("<a href='",str(?le_woewoe),"'>",?woewoe_form,"</a>") as ?woewoe)

 OPTIONAL
 { SELECT ?le_woewoe (GROUP_CONCAT(DISTINCT ?translation; separator="; ") as ?gloss)
   WHERE {
     ?le_woewoe ontolex:sense ?se_woewoe.
    [] vartrans:target ?se_woewoe;
       vartrans:source ?se_de.
    ?le_de ontolex:sense ?se_de;
           ontolex:canonicalForm/ontolex:writtenRep ?raw_translation.
     FILTER(lang(?raw_translation)='de')
     BIND(str(?raw_translation) as ?translation)
   } GROUP BY ?le_woewoe
 }


 OPTIONAL {
  SELECT ?le_woewoe (GROUP_CONCAT(?entry; separator="<br/>") as ?pdt)
  WHERE {
   []  vartrans:relates ?le_woewoe;
    vartrans:relates ?le_other;
    rdf:value ?y.
   FILTER(contains(str(?le_other),"plautdietsch"))
   ?le_other ontolex:canonicalForm/ontolex:writtenRep ?c.
            BIND(concat("<a href='",str(?le_other),"'>",?c,"</a> [",str(?y),"]") as ?entry)
  } GROUP BY ?le_woewoe
 }

 OPTIONAL {
  SELECT ?le_woewoe (GROUP_CONCAT(?entry; separator="<br/>") as ?plattmakers)
  WHERE {
   []  vartrans:relates ?le_woewoe;
    vartrans:relates ?le_other;
    rdf:value ?y.
   FILTER(contains(str(?le_other),"plattmakers"))
   ?le_other ontolex:canonicalForm/ontolex:writtenRep ?c.
            BIND(concat("<a href='",str(?le_other),"'>",?c,"</a> [",str(?y),"]") as ?entry)
  } GROUP BY ?le_woewoe
 }

 OPTIONAL {
  SELECT ?le_woewoe (GROUP_CONCAT(?entry; separator="<br/>") as ?reuter)
  WHERE {
   []  vartrans:relates ?le_woewoe;
    vartrans:relates ?le_other;
    rdf:value ?y.
   FILTER(contains(str(?le_other),"/dwn/"))
   ?le_other ontolex:canonicalForm/ontolex:writtenRep ?c.
            BIND(concat("<a href='",str(?le_other),"'>",?c,"</a> [",str(?y),"]") as ?entry)
  } GROUP BY ?le_woewoe
 }

 OPTIONAL {
  SELECT ?le_woewoe (GROUP_CONCAT(?entry; separator=" ") as ?twents)
  WHERE {
   []  vartrans:relates ?le_woewoe;
    vartrans:relates ?le_other;
```

```
   rdf:value ?y.
  FILTER(contains(str(?le_other),"twentswoordenboek"))
  ?le_other ontolex:canonicalForm/ontolex:writtenRep ?c.
           BIND(concat("<a href='",str(?le_other),"'>",?c,"</a> [",str(?y),"]") as ?entry)
 } GROUP BY ?le_woewoe
}

OPTIONAL {
 SELECT ?le_woewoe (GROUP_CONCAT(?entry; separator=" ") as ?wwb)
 WHERE {
  []  vartrans:relates ?le_woewoe;
   vartrans:relates ?le_other;
   rdf:value ?y.
  FILTER(contains(str(?le_other),"woerterbuchnetz.de/"))
  ?le_other ontolex:canonicalForm/ontolex:writtenRep ?c.
           BIND(concat("<a href='",str(?le_other),"'>",?c,"</a> [",str(?y),"]") as ?entry)
 } GROUP BY ?le_woewoe
}

FILTER(BOUND(?pdt) || BOUND(?plattmakers) || BOUND(?reuter) || BOUND(?twents) || BOUND(?wwb))
} ORDER BY ?woewoe ?le_woewoe
```