Bi-Text Mining Across German Dialects: On the Role of Synthetic Training Data for Dialect Adaptation

Jing Wang[1] **Barbara Plank**[1,2] **Robert Litschko**[1,2]
[1]MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
[2]Munich Center for Machine Learning (MCML), Munich, Germany
robert.litschko@lmu.de

Abstract

Cross-dialect bi-text mining relies on robust multilingual sentence representations to identify semantically equivalent sentence pairs across languages. While recent multilingual bi-encoder models achieve strong performance on standardized written languages, their behavior on dialectal varieties is largely unknown. In this study, we use Tatoeba to evaluate the performance of four widely-used bi-encoders on dialect-to-standard German translation retrieval, covering German documents and queries written in three dialects: Low German, Bavarian, and Alemannic. Motivated by the lack of resources, we examine the extent to which synthetic translations (from dictionaries and large language models; LLMs) can serve as weak supervision for dialect adaptation. Our results reveal that bi-encoders, when applied in a zero-shot setting, exhibit deficiencies in capturing semantic similarity between German and dialects, while fine-tuning on synthetic data substantially improves their retrieval effectiveness, with larger gains obtained from LLM-translated training data. We further analyze retrieval performance on Bavarian across varying dialect word proportions and observe a drop when dialect words make up more than 60% of the text.

Keywords: bi-text mining, synthetic data augmentation, dialect retrieval, German dialects.

1. Introduction

Most existing multilingual embedding models are optimized using large-scale parallel or comparable corpora involving standardized written languages, such as English, Chinese and Standard German (Zhang et al., 2024; Feng et al., 2022; Chen et al., 2024; Reimers and Gurevych, 2019, *inter alia*). These training regimes implicitly assume relatively stable orthography, consistent lexical conventions, and sufficient coverage across language varieties. Dialectal data, however, are often in strong contrast to these assumptions. Unlike many low-resource languages, dialects typically do not form independent standardized systems but exist in a continuum with the standard language, sharing large portions of vocabulary while simultaneously exhibiting significant culture-specific language use, such as regionally preferred lexical variants, local idioms, pragmatic particles, and discourse conventions. For example, the Standard German imperative phrase *"Beeil dich nicht"* (*"Don't hurry"*) may be realized in Swabian as *"No ned huddla"*,[1] and the first-person pronoun *ich* may appear as *i/ig* in Swiss German. Dialect-specific spelling variations and idioms reduce surface-level similarity with standard German and increase the lexical gap between languages (Berger et al., 2000). In this work, we study the alignment of text representations between standard German and dialects.

Large-scale dialect-Standard German parallel data is scarce, and dialects are largely absent from established bi-text mining benchmarks (e.g., prior BUCC shared tasks (Zweigenbaum et al., 2018, 2017)), which focus largely on standard languages. At the same time, dialect-aware machine translation evaluation (Deutsch et al., 2025; Vamvas et al., 2025) and representation learning (Philippy et al., 2025; Artemova and Plank, 2023) are becoming increasingly active research areas. However, their application on German dialects remains underexplored: beyond the Bavarian NMT case study of Her and Kruschwitz (2024), much of the prior work focuses on translation at the lexical level and the creation of dialect dictionaries (Bui et al., 2026; Chiarcos et al., 2025; Litschko et al., 2025a). Practically, this motivates dialect-standard German bi-text mining as a scalable way to extract aligned supervision from comparable sources for both evaluation and model adaptation. Moreover, recent work has demonstrated that synthetic query–document pairs generated by large language models (Jeronymo et al., 2023) or dictionary-based translations (Alam et al., 2024) can effectively augment semantically correct training data when authentic resource is limited. While these methods have shown promise for standard languages, their applicability to dialectal varieties remains uncertain.

Motivated by these gaps and trends, we study dialect-to-standard German retrieval. To this end, we create datasets from Tatoeba, using German documents and queries in Bavarian (bar), Low German (nds), and Alemannic (als). Our work is most similar to Artemova and Plank (2023), who compared how well the cosine similarity between Ger-

[1]The verb "huddla" is derived from the noun "huddel" and culturally grounded. See Appendix C for further information on the etymology of the word.

Proceedings of the 19th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC 2026, pages 72–83
11 May 2026.

man and Alemannic/Bavarian sentence embeddings aligns with human similarity judgments on a Likert scale. We evaluate bi-encoders on sentence-level retrieval and quantify performance gains obtained from fine-tuning on synthetic data. We also ablate their performance with respect to varying proportions of dialect terms. Taken together, we investigate bi-encoders under a more realistic and challenging evaluation protocol. In this work, we address the following research questions:

- **RQ1**: How well do state-of-the-art bi-encoders perform in bi-text mining when queries are written in German dialects, compared to when they are written in English?
- **RQ2**: To what extent does training on translated data from dictionaries and LLMs improve the retrieval performance of bi-encoders?
- **RQ3**: How robust is the performance of bi-encoders with respect to different ratios of dialect code mixing?

To summarize our contributions, we (1) propose an evaluation protocol for dialect-aware translation retrieval encompassing three German dialects; (2) provide a comprehensive evaluation of multilingual bi-encoders in both zero-shot and fine-tuned settings, offering insights into their strengths and limitations when applied to dialectal data; (3) we study synthetic data augmentation strategies for dialect retrieval and conduct ablation analyses on factors affecting task difficulty, while also discussing common quality issues in the dialect subsets of web-crawled datasets and their implications for evaluation reliability. Our code and data can be found at: https://github.com/mainlp/dialect-bitext-mining.

2. Related Work

NLP Research for German Dialects. Recent work in German dialect NLP spans resource building (Burghardt et al., 2016; Litschko et al., 2025a), annotation and dialect identification (Zampieri et al., 2017; Blaschke et al., 2024; Peng et al., 2024), information retrieval (Litschko et al., 2025b), and machine translation (Her and Kruschwitz, 2024; Aepli et al., 2023). In cross-dialect retrieval, Litschko et al. (2025b) explicitly cast dialect variation as an information access problem and introduce WikiDIR, a cross-dialect retrieval test collection derived from multiple German dialect Wikipedia. Though their framing aligns with our setting, there are key differences: we focus on bi-text mining with dense retrieval rather than keyword-oriented matching. Our study extends beyond evaluation and includes training of bi-encoders under synthetic supervision. We additionally study the robustness of bi-encoders under varying proportions of dialect tokens. Despite the growing interest in German dialect NLP, publicly available parallel corpora remain still limited. Blaschke et al. (2023) provide a systematic overview of resources for German dialects including corpora that contain dialect text aligned to Standard German (or multilingual) translations. In Section 3.2, we discuss quality aspects of parallel dialect data.

Dense Retrieval with Bi-Encoders. In dense retrieval, bi-encoders are used to independently embed queries and documents into continuous vector representations, which are then ranked based on similarity measures (e.g., cosine similarity) (Karpukhin et al., 2020; Reimers and Gurevych, 2019). Bi-encoders are trained using contrastive objectives that pull the embeddings of query-document pairs closer together while pushing apart those of queries and non-relevant documents (Karpukhin et al., 2020; Oord et al., 2019). Popular benchmarks for evaluating multilingual bi-encoders, such MMTEB (Enevoldsen et al., 2025), do not focus on dialectal variation, leaving the performance of the bi-encoders on dialects largely underexplored. To address this gap, we evaluate recent models on German dialects from the Tatoeba dataset (Tiedemann, 2020). Related to our work, Vamvas et al. (2024) investigate how continual pre-training of multilingual pre-trained language models affects sentence retrieval performance in Swiss German (gsw). The authors focus on unsupervised retrieval, where sentences are matched at the token-level using BERTScore (Zhang et al., 2020). In contrast, we study multilingual bi-encoders (single-vector retrieval) and evaluate their retrieval effectiveness after fine-tuning on synthetic training data.

Synthetic Data Augmentation. Since large-scale dialect-standard parallel data is scarce, a practical way to obtain supervision for training dense retrievers is to generate synthetic training data. Large language models (LLMs) have been shown to be effective synthetic data generators in information retrieval (IR) tasks (Thakur et al., 2024; Jeronymo et al., 2023; Harsha et al., 2025), as well as in machine translation (MT) for low-resource languages (Scalvini et al., 2025; de Gibert et al., 2025). In the context of MT, Kim et al. (2025) find that synthetic data generated by GPT-4o improves MT quality for low-resource languages. We extend this analysis to dialect bi-text mining and evaluate if GPT-4o translations improve the retrieval effectiveness of bi-encoders. Next to LLM-based generation, prior work has also explored dictionaries-based data augmentation for low-recourse MT (Alam et al., 2024; Nag et al., 2020). A key advantage of lexical trans-

	Tatoeba	WikiMatrix	Wikimedia	Total
nds-de	17,984	75,591	—	93,575
bar-de	90	41,991	3,351	45,432
als-de	1,714	—	1,149	2,863
de-eng	322,413	1,573,438	180,809	2,076,660

Table 1: The amount of available parallel sentences for each language pairs. Wikimedia statistics are based on the v20230407 release and the Swiss German (gsw) subdialect of Alemannic (als).

lation is that it allows us to precisely control the proportion of dialect words (Section 5.3).

3. Evaluation Protocol

3.1. Available Dialect Datasets

In our experiments, we pair Low German (nds), Bavarian (bar) and Alemannic (als) dialect queries with German (de) documents, yielding three language pairs {nds, bar, als}–de. We include two varieties of Alemannic: Swiss German and Swabian. These dialects are among the few varieties for which sentence-level parallel data with Standard German is publicly available in sufficient quantity to support systematic evaluation (Blaschke et al., 2023). Existing datasets vary in their coverage of our selected dialect pairs and number of instances. In this study, we use Tatoeba (Tiedemann, 2020), WikiMatrix (Schwenk et al., 2021) and Wikimedia, which have been made available by the OPUS project (Tiedemann, 2012).Table 1 provides an overview of each dataset and the number of available instances. Tatoeba is a crowd-sourced collection of user-provided translations, which has over 12.6M sentences in 426 languages and is widely used in low-resource and multilingual machine translation research. WikiMatrix is a large-scale automatically mined parallel corpus extracted from aligned Wikipedia articles using LASER (Artetxe and Schwenk, 2019), which contains 135M parallel sentences for 16,720 different language pairs in total (Schwenk et al., 2021). The extracted text is split into sentences and de-duplicated. Parallel texts in Wikimedia originate from Wikipedia articles that have been translated with computer-assisted translation tools and human oversight (Laxström et al., 2015).

3.2. Dataset Quality

According to Schwenk et al. (2021), the bi-text mining method adopted by WikiMatrix may lead to a drawback of increased misalignment risk, especially for low-resource languages. In particular, previous research on the quality of web-crawled multilingual datasets (Kreutzer et al., 2022) also shows that the ratio of correct samples from WikiMatrix is at a surprisingly low level. We assessed the data quality of the nds-de and bar-de pairs by examining 1,000 samples of each language pair. Our analysis revealed that a large proportion of the pairs is misaligned: 33.2% for nds-de and 47.7% for bar-de. We also find many instances where the Standard German sentence appears on the Bavarian side (28.9%). As an additional check, we compare how well translation pairs in WikiMatrix, Wikimedia, and Tatoeba can be retrieved using BM25 (Robertson and Zaragoza, 2009). Our results on WikiMatrix and Wikimedia (Table 2) are substantially higher than those on Tatoeba (Table 3), indicating a higher risk of introducing lexical shortcuts during evaluation. Taken together, our analyses reveal substantial lexical overlaps and misaligned pairs as factors that could bias the retrieval results. We therefore exclude both WikiMatrix and Wikimedia from the evaluation test set and proceed with Tatoeba (Sections 3.3).

3.3. Evaluation Data

Dialect-to-German Evaluation. We focus on the translation retrieval involving four language pairs {nds, als, bar, en}-de. Here, en-de serves as a reference point to compare the results against a high-resource language pair. Based on our analysis in Section 3.2, we select Tatoeba as a high quality dataset. In Tatoeba, texts consists of a mix of phrases and sentences. In the following, we refer to dialect translations as queries, and their German counterparts as documents. Following Litschko et al. (2019), we use 1K different queries for each dialect and 100K German documents. The document pool is shared between all language pairs and includes all 4K "relevant documents" (i.e., translations) and 96K randomly sampled nonrelevant documents. These negatives are Standard German texts sampled from Tatoeba's German-English subset. Given that Tatoeba contains only 90 Bavarian-German sentence pairs, we supplement the dataset with an additional 910 bar-de instances. These are generated by translating the German sentences from the German-English subset of Tatoeba into Bavarian. Models are evaluated using Mean Reciprocal Rank (MRR), Recall@10, and Precision@1.

	als-de (Wikimedia)			nds-de (Wikimedia)			bar-de (WikiMatrix)		
Model	MRR@10	R@10	P@1	MRR@10	R@10	P@1	MRR@10	R@10	P@1
BM25	0.451	0.537	0.410	0.221	0.334	0.175	**0.715**	**0.764**	**0.690**

Table 2: BM25 results of MRR@10, Recall@10, Precision@1 on Wikimedia bi-text for als-de and WikiMatrix sentence pairs for {nds, bar}-de. For als-de, we use 1K Wikimedia translation pairs of Swiss German and Standard German, and augment them with 99K German documents from WikiMatrix.

Dialect-Standard Mixtures. Dialects are frequently mixed with varying degrees of standard language terms. To investigate the retrieval performance with respect to different proportions of dialectal terms, we curate a dataset of 39 German sentences, each consisting of 10 words. These sentences are sampled from the German side of Tatoeba's English-German subset. We first tokenize each sentence into a 10-word list and prompt GPT-4o to generate a list of translations Bavarian. The model is constrained to output the same number of tokens (see Appendix A). Based on the word-aligned sentence pairs, we then separately substitute 20%, 40%, 60%, 80% and 100% of the original tokens in the German sentence with the translated Bavarian variants to generate 5 subsets with different portions of dialect words.

3.4. Synthetic Training Data

Weak Supervision. Motivated by the lack of large-scale training data for dialect bi-text mining, we evaluate two methods to obtain synthetic training data: dictionary-based word-by-word substitutions; LLM-based dialect translations. The synthetic data is not assumed to be fully correct or noise-free (Kim et al., 2025). Instead, it is used to simulate realistic low-resource training conditions, where manually curated parallel data is difficult to acquire on a large demand.

Dictionary-based Translations. In this approach, we use the Bavarian dialect variation dictionary (Litschko et al., 2025a) to generate synthetic training data through word-level code-switching. The dictionary is based on human annotations and provides Bavarian spelling variations for 5,124 German lemmas. To ensure a high vocabulary coverage, we use WikiMatrix as the source of parallel query-document pairs. We perform word-by-word substitution on both query and document sides: on the query side, we generate German-like documents by replacing dialect words (where available in the lexicon) with their Standard German equivalents; on the document side, we create dialect-like queries by substituting Standard German words with their Bavarian variants. This process expands each original de-bar WikiMatrix instance into multiple de-de$_{\text{bar}}$ and bar-bar$_{\text{de}}$ pairs. We limit the number of synthetic instances to at most 30 per sentence pair. The resulting dataset contains 32,458 instances with an average length of 26.4 tokens. We use our dictionary-based training data to evaluate whether models trained on Bavarian code-switched instances generalize to other dialects (**cross-dialect transfer**).

LLM-generated Translations. In this approach, we use GPT-4o-2024-08-06 (OpenAI, 2024) to translate Standard German sentences into dialects. We randomly select Standard German sentences from Tatoeba's German-English subset. To avoid data leakage, we ensured that none of the German sentences appear in any of the test splits. We then instruct the model to generate translations for each source-target language pair using the following prompt:

Translation Prompt

Translate the following Standard German sentence into natural, fluent `{target dialect}`. Only output the translation. Try to aim for diverse translations.

The output sentence is paired with the original Standard German sentence to form a synthetic parallel query-document pair. The synthetic subsets from each dialect–Standard German pair are merged into a single mixed parallel dataset (**multi-source training**). Based on prior work (Zhou et al., 2024; Lim et al., 2024), we expect that jointly training bi-encoders on multiple dialects will improve their performance on each individual dialect. The resulting dataset contains 27K translation pairs, with an average length of 7.7k tokens.

4. Experimental Setup

Models. In our experiments, we select four state-of-the-art multilingual sentence encoders that have been pretrained on large-scale cross-lingual data: LaBSE (Feng et al., 2022), gte-multilingual-base (Zhang et al., 2024), BGE-M3 (Chen et al., 2024) and Qwen3-Embedding-0.6B (Zhang et al., 2025). For retrieval, we rank documents according to their cosine similarity to the query. Although these models demonstrate strong retrieval performance

Model	als-de			nds-de			bar-de		
	MRR@10	R@10	P@1	MRR@10	R@10	P@1	MRR@10	R@10	P@1
Lexical Retrieval Baseline									
BM25	0.219	0.328	0.173	0.129	0.210	0.096	0.416	0.535	0.360
Zero-shot Evaluation									
LaBSE	**0.526**	**0.657**	**0.461**	**0.611**	**0.782**	**0.520**	**0.676**	**0.776**	**0.625**
GTE	0.427	0.554	0.363	0.532	0.701	0.449	0.592	0.693	0.54
BGE-M3	0.421	0.540	0.358	0.564	0.731	0.479	0.638	0.734	0.590
Qwen3	0.374	0.498	0.315	0.390	0.545	0.320	0.575	0.682	0.516
Avg.	0.437	0.563	0.374	0.524	0.690	0.442	0.620	0.721	0.570
Fine-tuning on dictionary-based translations									
LaBSE	**0.569**	**0.703**	**0.502**	0.656	0.815	0.570	0.692	0.790	0.639
GTE	0.510	0.650	0.444	0.617	0.794	0.526	0.659	0.754	0.616
BGE-M3	0.561	0.682	0.493	**0.711**	**0.849**	**0.637**	**0.726**	**0.819**	**0.677**
Qwen3	0.399	0.540	0.334	0.381	0.567	0.293	0.561	0.673	0.509
Avg.	0.510	0.644	0.443	0.591	0.756	0.507	0.659	0.759	0.610
Fine-tuning on LLM-generated translations									
LaBSE	0.811	0.898	0.762	0.853	0.956	0.784	0.896	0.962	0.860
GTE	0.793	0.891	0.729	0.851	0.960	0.775	0.889	0.958	0.850
BGE-M3	**0.849**	**0.921**	**0.797**	**0.880**	**0.969**	**0.822**	**0.936**	**0.981**	**0.908**
Qwen3	0.826	0.917	0.768	0.849	0.951	0.786	0.917	0.965	0.886
Avg.	0.820	0.907	0.764	0.858	0.959	0.791	0.909	0.967	0.876

Table 3: Results of evaluating bi-encoders without any training, and when trained with synthetic supervision. Results are reported in terms of Mean Reciprocal Rank at 10 (MRR@10), Recall at 10 (R@10) and Precision at 1 (P@1). We additionally report BM25 results (baseline). Best results are highlighted in **bold**.

on standard English–German setting, their performance on dialect-aware bi-text retrieval remains unclear. We include BM25 as a lexical baseline and as a proxy to measure task-level difficulty, providing a reference point for how much a task can be solved through lexical matching. We report our results using MRR@10, Recall@10 and Precision@1.

We evaluate bi-encoders in both zero-shot and fine-tuned settings. Training is based on Sentence-BERT (Reimers and Gurevych, 2019) using InfoNCE loss (Oord et al., 2019) with in-batch negatives. We select a batch size of 64 and maximum input sequence length 128 tokens. Both zero-shot evaluation and fine-tuning process are conducted on a single A100 GPU with 80 GB. On average, each run for training took 0.48 GPU-hours per model.

5. Results and Discussion

5.1. Zero-shot Evaluation

Table 3 (upper half) reports the results of our zero-shot retrieval experiments. Comparing the zero-shot retrieval results on dialects against those obtained on en-de Table 4 reveals a large gap. On average, bi-encoder models reach an MRR@10 reaches 0.861 on en-de. On dialect–Standard German language pairs, however, retrieval performance drops to MRR@10 scores ranging from 0.437 to 0.620. The lower zero-shot retrieval performance on dialect-Standard German pairs, compared to English-German (**RQ1**), highlights that current models are much better in aligning texts written in standard and high-resource languages. However, their performance deteriorates under dialect variation.

All dense retrieval models substantially outperform BM25, demonstrating the advantage of semantic matching over lexical overlap. Among them, LaBSE achieves the highest overall performance, leading in all three language pairs: it achieves MRR@10 of 0.526 for als-de, 0.611 for nds-de, and 0.676 for bar-de. The BGE-M3 and GTE models perform competitively, while Qwen3 performs the weakest, suggesting limited capacity to handle dialectal variation despite its strong multilingual foundation.

Among the dialect pairs, all models perform best on bar-de. BM25 achieves a performance of 0.416 MRR@10, while bi-encoders achieve a MRR@10 score of 0.721 on average (+0.305 MRR@10). On

Model	en-de		
	MRR@10	R@10	P@1
BM25	0.038	0.070	0.029
LaBSE	**0.918**	**0.990**	**0.867**
GTE	0.881	0.971	0.821
BGE-M3	0.901	0.978	0.850
Qwen3	0.806	0.938	0.721
Avg.	0.861	0.960	0.797

Table 4: Zero-shot retrieval results on the en-de portion of Tatoeba.

the other hand, the performance on als-de and nds-de is notably lower, with BM25 scores of 0.219 and 0.129, respectively, and bi-encoder results of 0.437 and 0.524, both below those for bar-de. This shows that lexical overlap directly relates to retrieval difficulty (**RQ3**). However, it is important to note that test instances bar-de consists mostly of synthetic instances. In Appendix B we quantify the evaluation gap between retrieval on authentic and synthetic data, and show that models obtain stronger performance on translated data.

5.2. Fine-tuning Evaluation

Table 3 (bottom half) shows our results obtained by fine-tuning bi-encoders on synthetic data. Across the board, fine-tuning on synthetic data yields substantial gains compared to zero-shot retrieval (**RQ2**). On average, fine-tuning on dictionary-based translations improves the zero-shot performance by +0.073 MRR@10 on als-de, +0.067 on nds-de, and +0.039 on bar-de. Fine-tuning on LLM-generated translations leads to much more pronounced improvements, with bi-encoders achieving MRR@10 gains of +0.383, +0.334, and +0.289 for the three dialects, respectively. These results are encouraging and show that weak supervision in the form of Bavarian code-switched data benefits the retrieval performance also on other dialects. The larger gains obtained with models trained on LLM-generated translations can likely be attributed to the fact that these translations provide full-text semantic equivalence, closely matching the test data. In contrast, dictionary-based synthesis is primarily word-level variant substitution rather than context-aware sentence translation. More crucially, the vocabulary coverage of our dictionary is limited, so that many content words in the synthetic dialect sentence remain unchanged Standard German words.[2] As a result, model training may collapse to capturing exact token matches.

[2]Our dictionary-based translation pairs still have a token overlap of 81.6%, while LLM-generated translation pairs show only 19.3% token overlap.

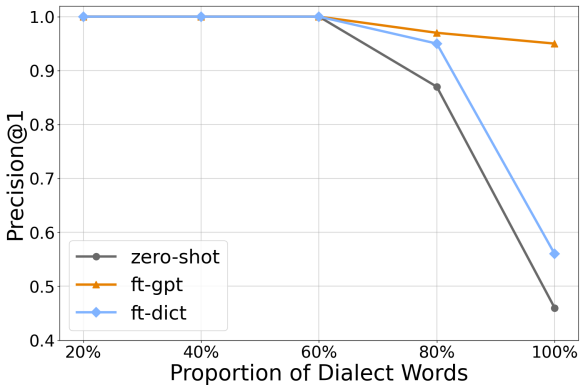

Figure 1: Results for LaBSE on bar-de with different ratios of Bavarian words. We compare zero-shot retrieval (zero-shot) to fine-tuning on LLM-translated (ft-gpt) and dictionary-translated data (ft-dict).

5.3. Robustness to Dialect Mixing

On the example of Bavarian, we now investigate how the dialect retrieval performance fluctuates with respect to different proportions of dialect tokens. In our experiments, we leave German documents unchanged and apply code switching on the query side. Figure 1 shows the retrieval results for LaBSE. We observe that retrieval performance remains consistently high when only 20%–60% of German tokens are replaced with their Bavarian translations. However, as the proportional of Bavarian words increases from 60%, performance deteriorates to varying extents. The zero-shot model exhibits the largest performance drop at 100% replacement, with P@1 reducing to 0.462. We also observe a sharp decline when LaBSE is fine-tuned on dictionary-translated data (0.564 P@1). In contrast, fine-tuning on LLM-generated data results in the most stable performance, with P@1 at 0.949. We find these trends to be consistent across bi-encoder models (see Appendix A). The results suggest that sufficient token overlap can compensate for a model's lack of dialect understanding (**RQ3**). Fine-tuning bi-encoders on LLM-generated data improves representation alignment and yields the most robust results.

6. Conclusion

This study evaluates four bi-encoder models for dialect-to-standard German retrieval. While all models outperform the lexical BM25 baseline, their retrieval performance lags behind when compared to English-to-German retrieval. We further show that fine-tuning on synthetic data consistently improves results, especially for LLM-generated translations. Our ablation on Bavarian-German reveals that the retrieval effectiveness starts to drop when the proportion of dialect words exceeds 60%.

7. Ethical considerations and limitations

Due to data scarcity, we did not evaluate model performance when trained on authentic dialect data. LLM-generated translations may introduce hallucinations or subtle meaning shifts (Vazquez et al., 2025), and dictionary-based substitutions often result in ungrammatical outputs and weak semantic equivalence (Alam et al., 2024). We quantify this difference in Appendix B.

Our focus lies on the alignment of dialect text representations with their corresponding Standard German translations. In practice, written dialects is used in social media, regional Wikipedia, and informal communication. Consequently, the scarcity of parallel data is reflective of the limited domains in which written dialects can be found. Future work should explore cross-modal alignment between dialects and Standard German in the speech domain.

Acknowledgements

We thank the anonymous reviewers for their invaluable feedback. This work is funded by the ERC Consolidator Grant DIALECT 101043235.

8. Bibliographical References

Noëmi Aepli, Chantal Amrhein, Florian Schottmann, and Rico Sennrich. 2023. A benchmark for evaluating machine translation metrics on dialects without standard orthography. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1045–1065, Singapore. Association for Computational Linguistics.

Md Mahfuz Ibn Alam, Sina Ahmadi, and Antonios Anastasopoulos. 2024. A morphologically-aware dictionary-based data augmentation technique for machine translation of under-represented languages.

Ekaterina Artemova and Barbara Plank. 2023. Low-resource bilingual dialect lexicon induction with large language models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 371–385, Tórshavn, Faroe Islands. University of Tartu Library.

Mikel Artetxe and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Adam L. Berger, Rich Caruana, David A. Cohn, Dayne Freitag, and Vibhu Mittal. 2000. Bridging the lexical chasm: statistical approaches to answer-finding. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Verena Blaschke, Barbara Kovačić, Siyao Peng, Hinrich Schütze, and Barbara Plank. 2024. MaiBaam: A multi-dialectal Bavarian Universal Dependency treebank. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10921–10938, Torino, Italia. ELRA and ICCL.

Verena Blaschke, Hinrich Schuetze, and Barbara Plank. 2023. A survey of corpora for Germanic low-resource languages and dialects. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 392–414, Tórshavn, Faroe Islands. University of Tartu Library.

Minh Duc Bui, Manuel Mager, Peter Herbert Kann, and Katharina von der Wense. 2026. Meenz bleibt meenz, but large language models do not speak its dialect.

Manuel Burghardt, Daniel Granvogl, and Christian Wolff. 2016. Creating a lexicon of Bavarian dialect by means of Facebook language data and crowdsourcing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2029–2033, Portorož, Slovenia. European Language Resources Association (ELRA).

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

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

Ona de Gibert, Joseph Attieh, Teemu Vahtola, Mikko Aulamo, Zihao Li, Raúl Vázquez, Tiancheng Hu, and Jörg Tiedemann. 2025. Scaling low-resource MT via synthetic data generation with LLMs. In *Proceedings of the*

2025 Conference on Empirical Methods in Natural Language Processing, pages 27674–27692, Suzhou, China. Association for Computational Linguistics.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects.

Duden. 2023. hudeln. Accessed: 2026-03-05.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, et al. 2025. Mmteb: Massive multilingual text embedding benchmark. In *The Thirteenth International Conference on Learning Representations*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding.

Hermann Fischer. 1911. *Schwäbisches Wörterbuch*, volume 3. Verlag der Laupp'schen Buchhandlung, Tübingen.

Chetan Harsha, Karmvir Singh Phogat, Sridhar Dasaratha, Sai Akhil Puranam, and Shashishekar Ramakrishna. 2025. Synthetic data generation using large language models for financial question answering. In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 76–95, Abu Dhabi, UAE. Association for Computational Linguistics.

Wan-hua Her and Udo Kruschwitz. 2024. Investigating neural machine translation for low-resource languages: Using Bavarian as a case study. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 155–167, Torino, Italia. ELRA and ICCL.

Vitor Jeronymo, Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, Jakub Zavrel, and Rodrigo Nogueira. 2023. Inpars-v2: Large language models as efficient dataset generators for information retrieval.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Seungone Kim, Juyoung Suk, Xiang Yue, Vijay Viswanathan, Seongyun Lee, Yizhong Wang, Kiril Gashteovski, Carolin Lawrence, Sean Welleck, and Graham Neubig. 2025. Evaluating language models as synthetic data generators. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6385–6403, Vienna, Austria. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Niklas Laxström, Pau Giner, and Santhosh Thottingal. 2015. Content translation: Computer-assisted translation tool for wikipedia articles. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*.

Seonghoon Lim, Taejun Yun, Jinhyeon Kim, Jihun Choi, and Taeuk Kim. 2024. Analysis of multi-source language training in cross-lingual transfer. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 712–725.

Robert Litschko, Verena Blaschke, Diana Burkhardt, Barbara Plank, and Diego Frassinelli. 2025a. Make every letter count: Building dialect variation dictionaries from monolingual corpora.

Robert Litschko, Goran Glavaš, Ivan Vulic, and Laura Dietz. 2019. Evaluating resource-lean cross-lingual embedding models in unsupervised retrieval. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1109–1112, New York, NY, USA. Association for Computing Machinery.

Robert Litschko, Oliver Kraus, Verena Blaschke, and Barbara Plank. 2025b. Cross-dialect information retrieval: Information access in low-resource and high-variance languages.

Ernst Martin and Hans Lienhart. 2012. *Wörterbuch der elsässischen Mundarten*. Walter de Gruyter.

Friedrich (ed.) Maurer, Friedrich Stroh, Rudolf Mulch, and Roland Mulch. 1973–1977. *Südhessisches Wörterbuch*, volume H–ksch of *Hessische Historische Kommission Darmstadt*. Marburg.

Josef Müller, Heinrich Dittmaier, Karl Meisen, and Matthias Zender. 1928–1971. Rheinisches wörterbuch, digitalisierte fassung im wörterbuchnetz des trier center for digital humanities.

Sreyashi Nag, Mihir Kale, Varun Lakshminarasimhan, and Swapnil Singhavi. 2020. Incorporating bilingual dictionaries for low resource semi-supervised neural machine translation.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding.

OpenAI. 2024. Gpt-4o system card.

Siyao Peng, Zihang Sun, Huangyan Shan, Marie Kolm, Verena Blaschke, Ekaterina Artemova, and Barbara Plank. 2024. Sebastian, Basti, Wastl?! recognizing named entities in Bavarian dialectal data. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14478–14493, Torino, Italia. ELRA and ICCL.

Fred Philippy, Siwen Guo, Jacques Klein, and Tegawende Bissyande. 2025. LuxEmbedder: A cross-lingual approach to enhanced Luxembourgish sentence embeddings. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11369–11379, Abu Dhabi, UAE. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*, volume 4. Now Publishers Inc.

Barbara Scalvini, Iben Nyholm Debess, Annika Simonsen, and Hafsteinn Einarsson. 2025. Rethinking low-resource MT: the surprising effectiveness of fine-tuned multilingual models in the LLM age. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 609–621, Tallinn, Estonia. University of Tartu Library.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Nandan Thakur, Jianmo Ni, Gustavo Hernandez Abrego, John Wieting, Jimmy Lin, and Daniel Cer. 2024. Leveraging LLMs for synthesizing training data across many languages in multilingual dense retrieval. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7699–7724, Mexico City, Mexico. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, volume 2012, pages 2214–2218.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Jannis Vamvas, Noëmi Aepli, and Rico Sennrich. 2024. Modular adaptation of multilingual encoders to written Swiss German dialect. In *Proceedings of the 1st Workshop on Modular and Open Multilingual NLP (MOOMIN 2024)*, pages 16–23, St Julians, Malta. Association for Computational Linguistics.

Jannis Vamvas, Ignacio Pérez Prat, Not Battesta Soliva, Sandra Baltermia-Guetg, Andrina Beeli, Simona Beeli, Madlaina Capeder, Laura Decurtins, Gian Peder Gregori, Flavia Hobi, Gabriela Holderegger, Arina Lazzarini, Viviana Lazzarini, Walter Rosselli, Bettina Vital, Anna

Rutkiewicz, and Rico Sennrich. 2025. Expanding the wmt24++ benchmark with rumantsch grischun, sursilvan, sutsilvan, surmiran, puter, and vallader.

Raul Vazquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sanchez Vega, Alessandro Raganato, Jindřich Libovický, Jussi Karlgren, Shaoxiong Ji, Jindřich Helcl, Liane Guillou, Ona De Gibert, Jaione Bengoetxea, Joseph Attieh, and Marianna Apidianaki. 2025. SemEval-2025 task 3: Mu-SHROOM, the multilingual shared-task on hallucinations and related observable overgeneration mistakes. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2472–2497, Vienna, Austria. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models.

Shijia Zhou, Huangyan Shan, Barbara Plank, and Robert Litschko. 2024. Mainlp at semeval-2024 task 1: Analyzing source language selection in cross-lingual textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1842–1853.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the Tenth Workshop on Building and Using Comparable Corpora*, pages 60–67, Vancouver, Canada. Association for Computational Linguistics.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

A. Dialect Mixing Ablation Study

For our ablation study (Section 5.3), we used GPT-4o to translate German sentences into Bavarian. We instructed the model to translate each given sentence word-by-word into Bavarian, following a structured output format. Input sentences are provided as a tokenized list of words.

Translation Prompt

You are translating German words into Bavarian.
Task:
Translate the following 10 German source words into exactly 10 Bavarian words, in the SAME order.
Rules:
- Return exactly 10 output words.
- Each German word must map to exactly ONE Bavarian word.
- No output word may contain spaces.
- Do not add punctuation. Do not change the order.
- Every output word must be different from the corresponding German source word (case-insensitive).
- Return ONLY a JSON object in this exact format:

```
{ "translations": ["w1", "w2",
"w3", "w4", "w5", "w6", "w7",
"w8", "w9", "w10"] }
```

Source words:
`{list of German words}`.
Only output the translation. Try to aim for diverse translations.

Figures 2 to 4 show the results for GTE, BGE-M3, and Qwen3 evaluated on varying proportions of dialect words. Overall, the trends are consistent with those reported for LaBSE (Section 5.3). That is, models demonstrate strong retrieval results if the proportion of dialect words is 60% or less.

B. Comparison Between Synthetic and Authentic Translations

In our main experiments evaluating Bavarian-standard German retrieval, we relied heavily on LLM-generated translations, with 910 out of 1,000

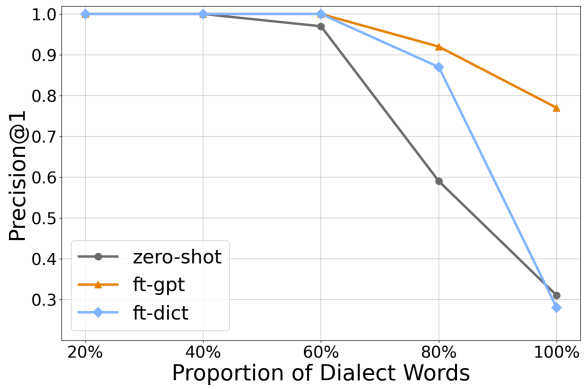

Figure 2: Results for GTE on bar-de with different ratios of Bavarian words. We compare zero-shot retrieval (zero-shot) to fine-tuning on LLM-translated (ft-gpt) and dictionary-translated data (ft-dict).

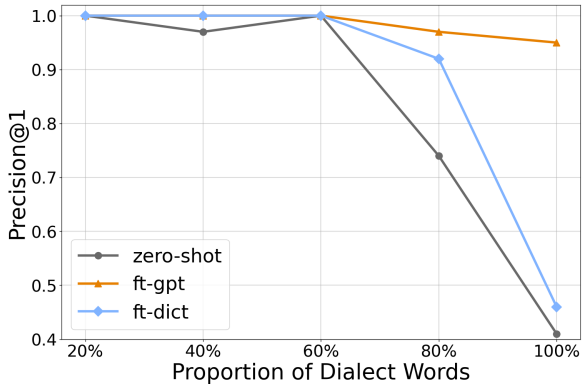

Figure 3: Results for M3 on bar-de with different ratios of Bavarian words. We compare zero-shot retrieval (zero-shot) to fine-tuning on LLM-translated (ft-gpt) and dictionary-translated data (ft-dict).

Bavarian queries being machine-translated (see Section 3.3). This was necessary due to the limited availability of human translations in Tatoeba, where we only had access to 90 authentic examples. However, this approach may introduce a bias, as bi-encoders trained on LLM-translated data may learn to recognize the characteristic "dialect style" of the GPT-4o model rather than genuine features of the Bavarian dialect. As a result, the reported performance gains may be inflated, and do not necessarily reflect the models' ability to capture authentic dialectal characteristics. To quantify this effect, we conduct an additional side-by-side comparison on the 90 authentic translation pairs, which we also translate using GPT-4o.

Table 5 shows the results of our best-performing bi-encoder (BGE-M3) evaluated on authentic translation pairs (top) and LLM-translated pairs (bottom). As expected, the model performs consistently better when evaluated on LLM-translated data. Fine-tuning and evaluating M3 on LLM-translated data

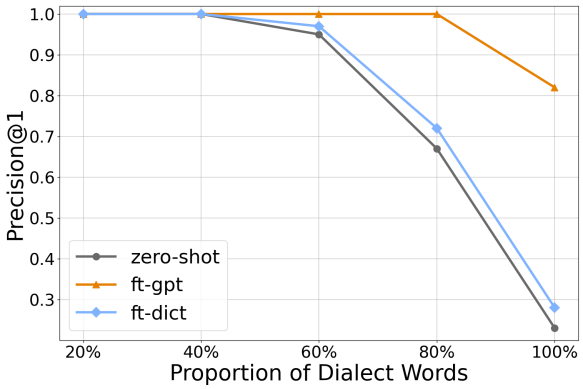

Figure 4: Results for Qwen3 on bar-de with different ratios of Bavarian words. We compare zero-shot retrieval (zero-shot) to fine-tuning on LLM-translated (ft-gpt) and dictionary-translated data (ft-dict).

Model	MRR@10	R@10	P@1
Zero-shot	0.288	0.425	0.219
Fine-tuned (LLM)	0.765	0.890	0.685
Fine-tuned (Dict)	0.406	0.548	0.329
Zero-shot	0.495	0.630	0.424
Fine-tuned (LLM)	0.867	0.987	0.781
Fine-tuned (Dict)	0.571	0.767	0.493

Table 5: Results for evaluating BGE-M3 on 90 parallel human-translated (top) and LLM-translated data (bottom).

yields the best results (0.781 P@1). Evaluating the same model on human translations reduces its retrieval effectiveness by 0.096 P@1 points. The performance gap is largest when the model is evaluated in a zero-shot fashion.

C. Explanation of "No ned huddla!"

The expression "No ned huddla!" ("nur nicht hudeln!") is taken from the Swabian Tatoeba dataset,[3] and means to not rush and be careless in a certain task (Duden, 2023). The Alemannic Wikipedia[4] traces the term "hudlâ" (dialect spelling variation) to traditional baking, where workers used damp cloths (="huddles") to quickly clean hot coals from ovens before baking bread, requiring swift action to prevent the cloth from burning. The etymological link is corroborated by documented regional dictionaries: the *Südhessisches Wörterbuch* (Maurer et al., 1973–1977, col. 760) records "hudeln" as "den angeheizten Backofen mit dem Huddel auswischen" (to wipe the heated oven with the

[3] https://tatoeba.org/en/sentences/show/6974751

[4] https://als.wikipedia.org/wiki/Wort:Schw%C3%A4bische_Vokabeln

Huddel); the *Rheinisches Wörterbuch* (Müller et al., 1928–1971, col. 885) defines “aushuddeln” as “den Backofen a., mit dem Huddel, dem Wischlumpen nach der Herausnahme der Kohlen auswischen” (to wipe out the oven with the rag after removing the coals); the *Wörterbuch der elsässischen Mundarten* (Martin and Lienhart, 2012, p. 304) lists “hudle” as “Den Backofen reinigen mit einem nassen Lumpen” (to clean the oven with a wet rag); and the *Schwäbisches Wörterbuch* (Fischer, 1911, p. 1851) lists the noun “hudel” as “Lumpen, mit dem der Bäcker den Backofen reinigt” (rags used by the baker to clean the oven).