Resource-Lean Lexicon Induction for German Dialects

Robert Litschko[1,2] **Barbara Plank**[1,2] **Diego Frassinelli**[1]
[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

Automatic induction of high-quality dictionaries is essential for building lexical resources, yet low-resource languages and dialects pose several challenges: limited access to annotators, high degree of spelling variations, and poor performance of large language models (LLMs). We empirically show that statistical models (random forests) trained on string similarity features are surprisingly effective for inducing German dialect lexicons. They outperform LLMs, enable cross-dialect transfer, and offer a lightweight data-driven alternative. We evaluate our models intrinsically on bilingual lexicon induction (BLI) and extrinsically on dialect information retrieval (IR). On BLI, random forests outperform Mistral-123b while being more resource-lean. On dialect IR with BM25, using our dialect dictionaries for query expansion yields relative improvements of up to 28.9% in nDCG@10 and 50.7% in Recall@100. Motivated by the resource scarcity in dialects, we further investigate the extent to which models transfer across different German dialects, and their performance under varying amounts of training data.

Keywords: Dialect variation, dictionary, cross-lingual transfer, German dialects, low-resource languages

1. Introduction

The performance of current natural language processing (NLP) tools and large language models (LLMs) crucially depends on the language of use and how well it is represented in the (pre-)training corpus. High-resource languages like English and German benefit from robust machine translation (MT) and mature information retrieval (IR) systems. However, their performance has been shown to deteriorate when translating non-standard languages (Gupta et al., 2025; Ziems et al., 2023) or retrieving relevant information where entities appear in regional spelling variations and word choices (Litschko et al., 2025b; Valentini et al., 2024; Chari et al., 2023). This leads to a lexical dialect gap between majority languages and dialects.

A major obstacle preventing the widespread adoption of NLP tools towards regional language varieties is the lack of resources and pretraining data, combined with the high variability due to the lack in standard orthography of many dialects. Multilingual language models cover only 1% of the world's over 7,000 languages in existence (Wang et al., 2022) and fail to accurately represent (i.e., tokenize) dialects at the input level (Muñoz-Ortiz et al., 2025; Srivastava and Chiang, 2025; Blaschke et al., 2023). Dictionaries offer a resource-lean alternative to capture dialect variation. For example in the context of retrieval, dictionaries have been traditionally used to translate queries to the document language (Ballesteros and Croft, 1996; Adriani and van Rijsbergen, 1999) and to induce cross-lingual embedding spaces (Litschko et al., 2018). In the context of translation, dictionaries have for example been used for data augmentation (Kale et al., 2020; Waldendorf et al., 2022; Yin et al., 2024), or to improve the ability of LLMs to translate texts into low-resource languages and rare words (Ghazvininejad et al., 2023; Lu et al., 2024). Compared to the (often unknown) language coverage of existing language models and translation systems, lexical resources exists for thousands of languages (Wang et al., 2022), with PanLex being the most prominent example (Baldwin et al., 2010; Kamholz et al., 2014). For dictionaries to be useful in practical applications they need to have a high lexical coverage (high recall), and high translation quality (high precision) to minimize erroneous translations. Meeting the coverage requirement is a particularly challenging task in the case of building dialect variation dictionaries, where words are mapped to multiple spelling variations. Manually curated dialect dictionaries developed by linguists achieve the highest quality standards, but are often unavailable in machine-readable form. We therefore focus on inducing bilingual dictionaries in a data-driven and lean way based on real-world dialect usage. Our method assumes to have access to Wikipedia for the target dialect, but is otherwise generalizable. To test the effectiveness of our approach, in this paper we specifically focus on five German dialects, each of which has its own Wikipedia: Alemmanic (als), Bavarian (bar), Ripuarian (ksh), Rhine Franconian (pfl), and Low German (nds).

Prior work on inducing German dialect dictionaries relied heavily on human supervision (Artemova and Plank, 2023; Litschko et al., 2025a,b). Artemova and Plank (2023) induce dictionaries based on word-alignments extracted from human-verified parallel sentences. This is not only costly and time-intensive due to the human supervision, but it also suffers from the same limitations as multilingual language models (i.e., most language varieties are not

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9044–9050
11–16 May 2026.

covered). Recent work tests LLMs for dictionary induction, but as shown in multiple studies (i.e., Li et al., 2023; Merx et al., 2024; Litschko et al., 2025a), building dictionaries with LLMs performs poorly in low-resource scenarios, especially for dialects. In contrast to these works, we show that statistical models trained on string similarity features are effective for building high-quality, high-coverage dictionaries at a much cheaper computational cost than LLMs, while achieving better overall quality. This opens up the support for more inclusive technology for minority languages. We further ablate the model performance with respect to the amount of training data used in the BLI task to determine how much data is truly needed to build high-quality resources. We make the following contributions:

- We create dialect variation dictionaries for five German dialects (§2.4).
- We validate the quality of our statistical model intrinsically on the task of BLI (§3.1), showing that it outperforms Mistral-123b.
- We evaluate our induced dictionaries extrinsically on the task of cross-dialect information retrieval (§3.3), showing consistent improvements across all dialects.

We make our code and resources available for future uptake.[1]

2. Methodology

2.1. Classifying Dialect Variation

We frame the task of bilingual induction as a word-pair classification task. The starting point for this is a vocabulary in German l_{de} (majority language) and in a regional language variety l_{dial} (dialect). The goal is to match terms in l_{de} to one or more terms in l_{dial}. Since the quadratic combination of all possible matches is too large to explore in practice, we use the DiaLemma annotation framework (Litschko et al., 2025a) to obtain for each term up to $k = 10$ lexical nearest neighbors, using Levenshtein distance (Levenshtein, 1966). This allows us to pre-filter promising candidate dialect words prior to scoring word-pairs based on more sophisticated string similarity features.

2.2. String Similarity Features

In this section, we describe the string similarity features used for our models, most of which have been used in Inkpen et al. (2005) in the context of classifying word pairs as cognates or false friends. The similarity measures are computed on pairs consisting of a German lemma x and a candidate dialect term y, and can be broadly grouped into set-based and sequence-based measures.

Set-based measures first transform the dialect term y and German term x into sets of ngrams. The first measure has been proposed by Adamson and Boreham (1974) and corresponds to the Dice-Sørensen coefficient computed on the shared character ngrams:

$$\textsc{Dice}(x, y) = \frac{2 \cdot |ngrams(x) \cap ngrams(y)|}{|ngrams(x) + ngrams(y)|} \quad (1)$$

Following Inkpen et al. (2005), we apply Dice on the set of bigrams, trigrams, and so-called "extended trigrams" consisting of trigrams minus their middle character (XDice; Brew et al., 1996). XXDice extends XDice by incorporating positional information, each overlapping token is weighted by

$$\frac{1}{1 + (pos(a) - pos(b))^2}, \quad (2)$$

where $pos(a)$ and $pos(b)$ correspond to the positions of the shared token in x and y. If a shared token appears multiple times we take the last occurrence. This scaling factor reduces the influence of matching ngrams that appear in different positions.

Sequence-based string similarity compare how well pairs of strings are aligned. The first feature simply measures the length the common prefix found in x and y (Prefix). The longest common subsequence ratio (LCSR; Melamed, 1999) relaxes the constraint that characters need to be adjacent, it counts the number of characters that appear in the same order divided by the length of the longer string. Following Inkpen et al. (2005), we also use the extended version of LCSR, which works on sequences of character bigrams (Bi-Sim) and trigrams (Tri-Sim) instead of sequences of individual characters (Kondrak and Dorr, 2004). In addition to sequence-based string similarity, we also use the edit distance between x and y normalized by the length of the longer sequence (NED). Here, too, we use the generalized version which computes NED on sequences of bigrams and trigrams (Inkpen et al., 2005; Kondrak and Dorr, 2004). Our final string similarity feature is based on phonetic similarity. We first encode both strings using the cologne phonetics algorithm (Postel, 1969) to transform both strings to their phonetic code, and then compute the edit distance between the two codes.

It is important to note that we deliberately use only features based on string similarity to evaluate how well pairs of German lemmas and dialect terms can be matched purely on surface-level features. The performance of our models can likely be improved by incorporating additional information such as term frequencies or part of speech categories.

[1] https://github.com/mainlp/dialect-lexicon-induction

2.3. Statistical Model

In this work, we use Random Forest (Breiman, 2001) implemented in the scikit-learn library (Pedregosa et al., 2011). We resort to the default values, where each random forest consists of 100 decision trees, using Gini impurity as a splitting criterion during training. In contrast to Litschko et al. (2025a), which applies logistic regression as a linear classification model, random forest allows for fitting non-linear decision boundaries. For each dialect and dataset (see §2.4), we train our model on 80% of the data, leaving 20% for testing. All reported results correspond to the average over repeated experiments with three different random seeds. A core advantage of statistical models is that they are much less resource-demanding and faster than LLMs, but also more effective in identifying dialect variations (§3).

2.4. Datasets and Evaluation

Bilingual Lexicon Induction (BLI). We evaluate our model intrinsically by measuring its performance on inducing bilingual dictionaries. Following prior work (Heyman et al., 2017; Irvine and Callison-Burch, 2017), we treat BLI as a classification task. Models are presented with a German word and a dialect candidate and must predict whether they represent translations of one another. Resorting to classification measures serves as a proxy to measure the dictionary quality in terms of its coverage (recall) and proportion of false entries (precision).

We evaluate on two recently published dialect variation dictionaries: DiaLemma (Litschko et al., 2025a) and WikiDIR (Litschko et al., 2025b). DiaLemma consists of 100k Bavarian word pairs in different word classes, while WikiDIR covers entities in five German dialects: Alemannic (als), Bavarian (bar), Low German (nds), Rhine Franconian (pfl), and Ripuarian (ksh). Both datasets are human-annotated, DiaLemma is uses three classes to indicate if a term is a dialect translation, an inflected variant, or unrelated, while WikiDIR adopts a binary label scheme. The datasets also differ in how candidate words were sourced: DiaLemma dictionaries are based on lexical nearest neighbors, while in WikiDIR they are derived from inter-language and inter-article links on Wikipedia.

Dialect retrieval. We evaluate our model extrinsically on the task of cross-dialect information retrieval. We specifically use the *analysis split*, where relevant documents contain query keywords in different spelling variations. For retrieval, we use the BM25 model (Robertson et al., 2009) implemented in the Pyserini library (Lin et al., 2021). Lexical retrieval models extract relevance signals from exact keyword matches and are ineffective when documents contain dialect variations of those keywords, since most dialects lack lexical normalization tools, such as stemmers and lemmatizers. We compare the performance of BM25 with the original queries against expanded queries, where we look up and append dialect spelling variations of query keywords. We measure the extent to which query expansion (QE) improves the results in terms of nDCG@10 and Recall@100.

For this experiment, we use the annotation framework proposed in Litschko et al. (2025a) for our five German dialects. For each dialect we 1) collect the 100K most frequent German lemmas, 2) find for each lemma its ten nearest lexical neighbors, and 3) classify word pairs whether they correspond to translations. The classification model used for step 3) is trained on the full DiaLemma dictionary (5.2K lemmas). We construct our dialect dictionaries by including all word pairs that the model identified as translation equivalents. The resulting dataset statistics are shown in Table 1.

Dialect	Lemmas	Variants	V/L
als	38,129	88,114	2.31
bar	27,598	51,392	1.86
ksh	6,889	9,384	1.36
pfl	9,127	13,050	1.43
nds	21,974	39,547	1.80

Table 1: Dataset statistics of our induced dictionaries. We show the number of lemmas for which we found at least one dialect variant, the total number of dialect terms, and the average per lemma.

Model	P	R	F1
Random	0.112	0.341	0.169
Mistral-123b	0.443	**0.743**	0.555
Random Forest	**0.646**	0.534	**0.585**

Table 2: Comparison between our Random Forest model against the Litschko et al. (2025a) best large language model (Mistral-123b) and a random baseline on Bavarian.

3. Results and Discussion

To evaluate the effectiveness of the Random Forest classifier in judging translation candidates, we start by comparing its performance against the overall results reported in Litschko et al. (2025a, Table 3). Using the DiaLemma dictionary, we contrast our model with the predictions for Mistral 123b (Jiang et al., 2023), their best performing LLM, and a random baseline. Table 2 shows a strong improvement against the random baseline, and it shows that a simpler and faster statistical model can outperform

train / test	ksh	nds	als	pfl	bar
ksh	**0.79**	0.62	0.77	**0.82**	0.67
nds	0.60	0.68	0.63	0.61	0.63
als	0.75	**0.71**	**0.80**	0.81	**0.70**
pfl	**0.79**	0.61	0.76	**0.82**	0.67
bar	0.72	0.67	0.72	0.70	0.68
ALL	0.77	**0.71**	0.80	0.81	**0.70**

Table 3: BLI - Cross-dialect $\mathbf{F_1}$ scores, with models trained on a specific source dialect and tested on all five target dialects. The highest score for each dialect is shown in bold.

train / test	ksh	nds	als	pfl	bar
ksh	0.76	0.49	0.67	0.77	0.57
nds	**0.93**	**0.77**	**0.80**	**0.94**	**0.75**
als	0.84	0.66	0.78	0.87	0.67
pfl	0.78	0.49	0.68	0.80	0.58
bar	0.83	0.66	0.78	0.84	0.70
ALL	0.85	0.67	0.79	0.90	0.67

Table 4: BLI - Cross-dialect **precision** scores, with models trained on a specific source dialect and tested on all five dialects. The highest score for each dialect is shown in bold.

train / test	ksh	nds	als	pfl	bar
ksh	**0.81**	**0.85**	**0.89**	**0.87**	**0.82**
nds	0.45	0.60	0.53	0.45	0.54
als	0.67	0.77	0.82	0.76	0.75
pfl	0.79	0.80	0.86	0.86	0.79
bar	0.64	0.69	0.67	0.59	0.66
ALL	0.71	0.76	0.81	0.74	0.73

Table 5: BLI - Cross-dialect **recall** scores, with models trained on a specific source dialect and tested on all five target dialects. The highest score for each dialect is shown in bold.

the best LLM used in previous work. In the following sections, we therefore report both intrinsic and extrinsic evaluations of the Random Forest model.

3.1. Bilingual Lexicon Induction

Table 3 reports the F_1 scores for BLI across the five dialects in WikiDIR. The last row shows the results of a multi-source model that is trained on the concatenation of all training splits. Overall, the model achieves a good performance ($F_1 = 0.75 \pm 0.07$) when trained and tested on the same dialect (i.e., the values on the diagonal). This confirms that the model can capture dialect-specific regularity in an effective way. When transferring across dialects, we observe varying results. When training on a single dialect, Alemannic shows the best results across all dialects (als row, $F_1 = 0.75 \pm 0.05$). In contrast, models trained on Low German data show a lower transferability (nds row, $F_1 = 0.63 \pm 0.03$). Across the board, combining the training data of all languages (ALL) achieves the highest F_1 scores across $F_1 = 0.76 \pm 0.05$), indicating that the diversity and volume of training data positively impact cross-dialect transfer.

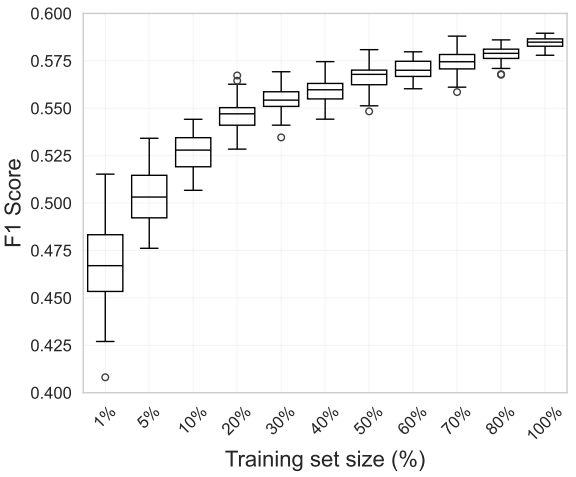

Figure 1: Effect of training set size on F1 scores for DiaLemma word pairs. We repeated our experiments with 40 different random seeds and evaluated on a fixed 20% test split.

Tables 4 and 5 show that Low German (nds) and Ripuarian (ksh) stand out as the best source languages, yielding the highest precision and recall values across all dialects. This suggests that the training data of Low German has a more consistent or less varied feature distribution, while the training data of Ripuarian might capture a more diverse set of features. Further research is needed to investigate the factors that influence cross-dialectal transfer at the lexical level.

3.2. Effect of Training Size

Figure 1 shows the effect of training size on model performance when trained on different portions of the training data and evaluated on the same 20% test split. We use word pairs from the DiaLemma dictionary, where full training and test datasets contain 80k and 20k training instances. We find that using only 10% of the training data (8k word pairs) yields an average F_1 score of 0.52, and training on 40% of the data ($F_1 = 0.56$) outperforms Mistral (see Table 2). A further increase in training size leads to a modest improvement of 5% ($F_1 = 0.59$). This indicates that even a relatively small training set is sufficient to obtain a strong performance, which is promising for dialects where data is scarce.

	nDCG@10				Recall@100				Statistics		
	BM25	**QE**	**Δ**	**Δ%**	**BM25**	**QE**	**Δ**	**Δ%**	**n_aug**	**n_query**	**% n_aug**
ksh	0.33	0.38	0.06	18.2%	0.31	0.36	0.06	18.0%	135	210	64.3%
nds	0.35	0.39	0.04	12.1%	0.28	0.34	0.06	20.0%	270	470	57.4%
als	0.36	0.39	0.03	8.7%	0.34	0.43	0.09	27.5%	2021	4639	43.6%
pfl	0.28	0.36	0.08	28.9%	0.21	0.31	0.11	50.7%	76	157	48.4%
bar	0.45	0.48	0.03	5.7%	0.41	0.49	0.08	18.9%	298	718	41.5%
ALL	0.35	0.40	0.05	14.71%	0.33	0.39	0.08	25.1%	560	1239	51.0%

Table 6: Cross-Dialect Information Retrieval - We show for each dialect the result of applying BM25 on the original queries (BM25) and after query expansion (QE), as well as the absolute (Δ) and relative differences (Δ%) in retrieval performance. **n_aug** and **%n_aug** refer to the absolute and relative number of augmented queries. **n_query** denotes the total number of queries.

3.3. Cross-Dialect Information Retrieval

Table 6 reports the results of the cross-dialect retrieval experiments using BM25 as baseline and a query-expanded (QE) variant across five dialects. As indicated by the positive deltas, query expansion consistently improves the retrieval results of BM25. On average, we observe an improvement (relative improvement) of +0.05 nDCG@10 (+14.71%) and +0.08 Recall@100 (+25.1%). This improvement is particularly evident for Rhine Franconian (pfl), which is the dialect with the lowest overall number of queries and the smallest document corpus. However, despite its smaller scale, BM25 (without QE) achieves the lowest performance on Rhine Franconian. In terms of coverage, we observe that approximately 51% of all German queries included at least one keyword for which dialect spelling variations are available in our dictionaries.

4. Conclusion

In this work, we show that statistical models trained on elaborate string similarity features are not only more resource-lean, both in terms of (pre-)training requirements and at inference time, but also more effective. We contribute dialect variation dictionaries for five German dialects, covering more dialects than DiaLemma (5 vs. 1), and substantially more lemmas than WikiDIR (103,717 vs. 6,257). In our extrinsic evaluation, we show that query expansion with our dictionaries consistently improves the dialect retrieval performance with BM25.

5. Limitations

In this work, we focus exclusively on surface-level features and do not incorporate (contextual) semantic similarity. While this has the advantage of being resource-efficient, it does not account for ambiguous terms. However, this affects only a minority of the cases. In future work, we plan to jointly model orthographic and semantic features.

6. Ethical Considerations

We see no ethical issues related to this work. All experiments were conducted with publicly available data and open-source software, and our code and linguistic resources are openly available on GitHub.

Use of AI Assistants The authors acknowledge the use of ChatGPT for correcting grammatical errors and enhancing the coherence of the final manuscript.

7. Bibliographical References

George W Adamson and Jillian Boreham. 1974. The use of an association measure based on character structure to identify semantically related pairs of words and document titles. *Information storage and retrieval*, 10(7-8):253–260.

Mirna Adriani and C. J. van Rijsbergen. 1999. Term similarity-based query expansion for cross-language information retrieval. In *Research and Advanced Technology for Digital Libraries*, pages 311–322, Berlin, Heidelberg. Springer Berlin Heidelberg.

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.

Timothy Baldwin, Jonathan Pool, and Susan Colowick. 2010. PanLex and LEXTRACT: Translating all words of all languages of the world. In *Coling 2010: Demonstrations*, pages 37–40, Beijing, China. Coling 2010 Organizing Committee.

Lisa Ballesteros and Bruce Croft. 1996. Dictionary methods for cross-lingual information retrieval. In *Database and Expert Systems Ap-*

plications, pages 791–801, Berlin, Heidelberg. Springer Berlin Heidelberg.

Verena Blaschke, Hinrich Schütze, and Barbara Plank. 2023. Does manipulating tokenization aid cross-lingual transfer? a study on POS tagging for non-standardized languages. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 40–54, Dubrovnik, Croatia. Association for Computational Linguistics.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Chris Brew, David McKelvie, et al. 1996. Word-pair extraction for lexicography. In *Proceedings of the 2nd international conference on new methods in language processing*, pages 45–55. Citeseer.

Andreas Chari, Sean MacAvaney, and Iadh Ounis. 2023. On the effects of regional spelling conventions in retrieval models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2220–2224.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *arXiv preprint arXiv:2302.07856*.

Abhay Gupta, Jacob Cheung, Philip Meng, Shayan Sayyed, Austen Liao, Kevin Zhu, and Sean O'Brien. 2025. Endive: A cross-dialect benchmark for fairness and performance in large language models. *arXiv preprint arXiv:2504.07100*.

Geert Heyman, Ivan Vulić, and Marie-Francine Moens. 2017. Bilingual lexicon induction by learning to combine word-level and character-level representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1085–1095, Valencia, Spain. Association for Computational Linguistics.

Diana Inkpen, Oana Frunza, and Grzegorz Kondrak. 2005. Automatic identification of cognates and false friends in French and English. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, volume 9, pages 251–257.

Ann Irvine and Chris Callison-Burch. 2017. A comprehensive analysis of bilingual lexicon induction. *Computational Linguistics*, 43(2):273–310.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Mihir Kale, Sreyashi Nag, Varun Lakshinarasimhan, and Swapnil Singhavi. 2020. Incorporating bilingual dictionaries for low resource semi-supervised neural machine translation. In *International Conference on Learning Representations, Learning with Limited Labeled Data*.

David Kamholz, Jonathan Pool, and Susan M Colowick. 2014. Panlex: Building a resource for panlingual lexical translation. In *LREC*, volume 14, pages 3145–3150.

Grzegorz Kondrak and Bonnie Dorr. 2004. Identification of confusable drug names: A new approach and evaluation methodology. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 952–958, Geneva, Switzerland. COLING.

Vladimir Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710. [Russian original (1965) in Doklady Akademii Nauk SSSR, 163(4):845—848.].

Yaoyiran Li, Anna Korhonen, and Ivan Vulić. 2023. On bilingual lexicon induction with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9577–9599, Singapore. Association for Computational Linguistics.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2356–2362.

Robert Litschko, Verena Blaschke, Diana Burkhardt, Barbara Plank, and Diego Frassinelli. 2025a. Make every letter count: Building dialect variation dictionaries from monolingual corpora. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, China. Association for Computational Linguistics.

Robert Litschko, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2018. Unsupervised cross-lingual information retrieval using monolingual data only. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1253–1256.

Robert Litschko, Oliver Kraus, Verena Blaschke, and Barbara Plank. 2025b. Cross-dialect information retrieval: Information access in low-resource and high-variance languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10158–10171, Abu Dhabi, UAE. Association for Computational Linguistics.

Hongyuan Lu, Haoran Yang, Haoyang Huang, Dongdong Zhang, Wai Lam, and Furu Wei. 2024. Chain-of-dictionary prompting elicits translation in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 958–976, Miami, Florida, USA. Association for Computational Linguistics.

I. Dan Melamed. 1999. Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25(1):107–130.

Raphael Merx, Ekaterina Vylomova, and Kemal Kurniawan. 2024. Generating bilingual example sentences with large language models as lexicography assistants. In *Proceedings of the 22nd Annual Workshop of the Australasian Language Technology Association*, pages 64–74, Canberra, Australia. Association for Computational Linguistics.

Alberto Muñoz-Ortiz, Verena Blaschke, and Barbara Plank. 2025. Evaluating pixel language models on non-standardized languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6412–6419, Abu Dhabi, UAE. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Hans Joachim Postel. 1969. Die kölner phonetik. ein verfahren zur identifizierung von personennamen auf der grundlage der gestaltanalyse. *IBM-Nachrichten*, 19:925–931.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Aarohi Srivastava and David Chiang. 2025. We're calling an intervention: Exploring fundamental hurdles in adapting language models to nonstandard text. In *Proceedings of the Tenth Workshop on Noisy and User-generated Text*, pages 45–56, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Francisco Valentini, Viviana Cotik, Damián Furman, Ivan Bercovich, Edgar Altszyler, and Juan Manuel Pérez. 2024. Messirve: A large-scale spanish information retrieval dataset. *arXiv preprint arXiv:2409.05994*.

Jonas Waldendorf, Alexandra Birch, Barry Hadow, and Antonio Valerio Micele Barone. 2022. Improving translation of out of vocabulary words using bilingual lexicon induction in low-resource machine translation. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 144–156.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 863–877.

Yongjing Yin, Jiali Zeng, Yafu Li, Fandong Meng, and Yue Zhang. 2024. LexMatcher: Dictionary-centric data curation for LLM-based machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14767–14779, Miami, Florida, USA. Association for Computational Linguistics.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-VALUE: A framework for cross-dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768, Toronto, Canada. Association for Computational Linguistics.