Model Selection for Homophone Spelling Correction in Low-Resource Languages: A Systematic Review with Khmer Case Study

Seanghort Born[1,2*] , Madeth May[1], Claudine Piau-Toffolon[1], and Sébastien Iksal[1]

[1]LIUM, Le Mans University, France

[2] Institute of Digital Research and Innovation, Cambodia Academy of Digital Technology, Cambodia

*Corresponding author
doi: https://doi.org/10.21467/proceedings.8.1.2

Abstract

This study conducts a comprehensive review of existing literature on spelling correction models designed for low-resource languages (LRLs). The research examines how these models handle homophone errors, using Khmer as a case study to highlight challenges in LRLs. Out of 174 academic publications, 54 papers were chosen for detailed study. The review covers various spelling correction models, including traditional, deep learning, and large language models. Traditional methods are affordable to implement but struggle to understand word context properly. Deep learning models (DLMs) provide the best balance between cost and effectiveness for correcting Khmer homophones. Large language models (LLMs) offer the highest accuracy but require significant computational power and may favor languages with more available data. The study also provides a three-stage framework for selecting appropriate models. When data is limited, traditional methods work best for homophone correction. When moderate amounts of data are available, DLMs should be the preferred choice. When both sufficient data and computational resources are accessible, LLMs deliver optimal results. These recommendations offer valuable guidance for researchers who develop spelling correction systems for LRLs, particularly when addressing homophone-related challenges.

Keywords: spelling correction, model selection, low-resource languages

Introduction

Spelling correction, especially for homophones, is vital in natural language processing (NLP), particularly for low-resource languages (LRLs). LRLs lack sufficient data for NLP tasks, mainly due to small speaker populations, weak infrastructure, and limited research investment (Eskander et al., 2019; Z. Liu et al., 2022). A key challenge for these languages is the lack of a standardized approach to model selection.

This study addresses these challenges through a systematic literature review (SLR), proposing a structured methodology for selecting language models for spelling correction. The objective is to evaluate existing models, focusing on their effectiveness in homophone detection and correction in LRLs, with Khmer as a case study. It examines the challenges, methodologies, and key findings from the literature, offering valuable insights for model selection in LRLs. Additionally, it provides practical guidelines for researchers to choose suitable models for homophone spelling correction (HSC) in resource-constrained environments. The next section reviews prior work on spelling correction, including studies on LRLs, Khmer, and homophones, highlighting current gaps in the field.

Proceedings DOI: 10.21467/proceedings.8.1; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-989164-6-4

The Current Spelling Correction Models

Automatic spelling error correction has been a major NLP research area for decades, with models designed to detect and fix errors caused by cognitive or typing mistakes (Cissé & Sadat, 2023; Gueddah & Lachibi, 2023). A wide range of models have been developed, from traditional approaches to advanced deep learning and large language models. Traditional spelling correction models provided the groundwork for current systems, comprising a dictionary, error model, and language model (Attia et al., 2016). Early methods relied on edit operations and string distance metrics such as Levenshtein and Damerau-Levenshtein (Eger et al., 2016; Gueddah et al., 2022; Mukazhanov et al., 2023), along with confusion matrices for character-level error estimation (Hladek et al., 2020; Mammadov, 2019). Rule-based techniques enabled language-specific string matching (Dong et al., 2019; Gueddah & Lachibi, 2023), while probabilistic and n-gram models improved error prediction and candidate ranking (Gueddah et al., 2022; Lai et al., 2015). The noisy channel model, enhanced by contributions from (Attia et al., 2016; Hasan et al., 2015), combined error and language models using Bayes' theorem (Kim et al., 2022; Mainsah et al., 2015). Dictionary- and frequency-based approaches further aided in detecting and prioritizing corrections (Kasmaiee et al., 2023; Lai et al., 2015). These models laid the foundation for modern systems, improving error correction in simple, context-free spelling tasks. In contrast, deep learning expanded these foundations by enabling neural models to learn error patterns and context (Hladek et al., 2020). Bi-LSTM and LSTM proved effective across various languages (Born et al., 2022; Kusuma & Ratnasari, 2023; Moslem et al., 2020). Sequence-to-Sequence (Seq2Seq) models mapped misspellings to corrections directly (Etoori et al., 2018; Salhab & Abu-Khzam, 2024; S. Zhang et al., 2020), while Transformer-based models improved performance with attention mechanisms (Do et al., 2021; Lertpiya et al., 2020). BERT-based models incorporated phonetic and contextual features for better accuracy (Liang et al., 2023; R. Zhang et al., 2021). Character-level CNNs and RNNs handled out-of-vocabulary errors (Kim et al., 2022; Kuznetsov & Urdiales, 2021), and hybrid approaches combined traditional and neural methods for enhanced performance (Kasmaiee et al., 2023). Despite their data needs, neural models are robust across languages and error types (Hladek et al., 2020; Mammadov, 2019). Recently, like BERT and T5 have transformed spelling correction by leveraging deep contextual understanding (S. Liu et al., 2021; Moslem et al., 2020). Through masked language modeling and Seq2Seq architectures, LLMs effectively detect and correct errors (S. Liu et al., 2021; Tran et al., 2021; Zhou et al., 2017). They also support zero- and few-shot corrections (Sun et al., 2023), and integrating them with traditional methods further boosts accuracy (Hu et al., 2020), offering highly adaptable, context-aware correction systems. In conclusion, traditional, deep learning, and large language models have each advanced spelling correction. Traditional models introduced basic error modeling and context use. Deep learning improved the handling of complex patterns. LLMs added stronger contextual understanding and flexibility. These advancements have made spelling correction more accurate and adaptable across languages.

Spelling Correction in LRLs

Spelling correction is a critical task in NLP, aimed at detecting and rectifying text errors (Do et al., 2021; Hladek et al., 2020; Mammadov, 2019). While it has been extensively studied for high-resource languages (HRLs), it presents unique challenges for LRLs (Büyük & Arslan, 2021; Goslin & Hofmann, 2022; Magueresse et al., 2020). These challenges arise mainly from the lack of large, clean

Proceedings DOI: 10.21467/proceedings.8.1
ISBN: 978-81-989164-6-4

Series: AIJR Proceedings
ISSN: 2582-3922

training datasets and specific linguistic resources (Büyük & Arslan, 2021; Etoori et al., 2018; Magueresse et al., 2020). Various approaches have been explored to address spelling correction in LRLs, such as rule-based systems for Marathi, Persian, and Kazakh, which utilize morphological, orthographical, phonetic rules, and finite state automata (Etoori et al., 2018; Kasmaiee et al., 2023; Mukazhanov et al., 2023). In addition, n-gram models have been applied to real-word error correction in Bangla and spelling correction in Tibetan (Mashod Rana et al., 2018; San et al., 2021). Moreover, deep learning models (DLMs) have been explored for spelling correction in languages like Hindi and Telugu, Vietnamese, and Azerbaijani (Etoori et al., 2018; Mammadov, 2019; Tran et al., 2021). In Khmer language, solutions include the Khmer Spelling Checker (KSC) for Microsoft Word, word segmentation using bi-directional maximal matching, and neural approaches like RNNs and LSTMs for historical text correction (Born et al., 2022; Mao et al., 2022; Van Nam et al., 2017). Traditional dictionary- and rule-based methods also play a vital role in spelling correction in this LRL (Sung & Hwang, 2016; Van Nam et al., 2017).

These efforts illustrate the diverse methods used to address spelling correction in LRLs.

In summary, spelling correction in LRLs remains a challenging but vital NLP task. Despite progress with rule-based, statistical, and neural methods, each language poses unique challenges. More language-specific solutions are needed, supported by ongoing data collection, model development, and linguistic analysis.

Challenges of Homophones in NLP Systems

Homophone errors present significant challenges for authors across many languages due to their similar pronunciations, leading to frequent mistakes in writing (Born et al., 2024). These errors are not limited to typographical issues, such as incorrect keystrokes (Cissé & Sadat, 2023; Mammadov, 2019), but are often cognitive in nature, arising from uncertainty about the correct spelling for the intended meaning (Born et al., 2024; Cissé & Sadat, 2023; Mammadov, 2019). Unlike non-word errors that can be detected through dictionary lookup (Fahrudin et al., 2021; Lai et al., 2015; Phung & Luong, 2024), homophone errors occur in sentences where each word appears correctly spelled, yet the overall meaning is flawed (Born et al., 2024). These threats make detection and correction significantly more complex. For instance, Chinese faces significant challenges with homophonic errors, with research showing that about 83% of spelling mistakes in Chinese are due to the misuse of phonologically similar characters (homophones) (Liang et al., 2023; Lin et al., 2024; L. Liu et al., 2024). These errors are difficult to correct, especially in short sentences that lack sufficient semantic context (Liang et al., 2023). In English, while homophones are acknowledged as a challenge, the research does not delve deeply into specific homophone correction issues (Hu et al., 2020). Khmer, Cambodia's official language and a LRL, faces notable spelling challenges due to its complex script and homophones (Buoy et al., 2023). A recent study highlighted the lack of a HSC model for Khmer, emphasizing a critical research gap (Born et al., 2024). While HRLs benefit from DLMs trained on large datasets, Khmer lacks sufficient linguistic data. This calls for alternatives like data augmentation, transfer learning, and hybrid rule-based models. Still, using context to distinguish homophones remains essential across all languages.

Proceedings DOI: 10.21467/proceedings.8.1
ISBN: 978-81-989164-6-4

Series: AIJR Proceedings
ISSN: 2582-3922

To conclude, homophones pose both linguistic and computational challenges for NLP. Khmer, with its complex script and homophones, highlights these difficulties.

Although progress exists, more research is needed to address Khmer's unique traits. In LRLs like Khmer, this demands focused effort. Strong correction techniques are vital for improving NLP accuracy, accessibility, and language learning.

Methodology

To support the development of an effective HSC model for LRLs, this study followed a structured, multi-phase approach. It began with a systematic literature review to identify and analyze relevant research on spelling correction, homophones, and NLP techniques applicable to LRLs. The findings informed a comparative analysis of existing models and a conceptual case study focused on the Khmer language, serving as a representative example of LRL challenges. Based on this foundation, development guidelines were formulated to address the specific needs of homophone correction in resource-constrained settings.

The Paper Selection Process

A systematic literature review (SLR) was conducted to gather peer-reviewed publications from prominent academic databases including IEEE Xplore, ACL Anthology, Google Scholar, Semantic Scholar, ScienceDirect, and the Association for Computing Machinery (ACM). These databases were selected because they provide extensive coverage of natural language processing, machine learning, and linguistics research. The search strategy employed various keywords such as "automatic spelling correction," "misspelling correction," "homophones," "Khmer homophones," "homophone spelling correction," "natural language processing," "low-resource languages," and "Khmer language." These terms were used both individually and in combination to improve search results. After searching these databases and removing duplicate articles, the researchers identified 174 potentially relevant papers, consisting of 112 conference papers and 62 journal articles. To ensure currency and relevance, 31 publications from before 2015 were excluded, focusing on research from the most recent decade (2015-2025). The remaining 143 papers underwent initial screening through title and abstract review, with selection based on four key criteria: spelling correction research, homophone-related studies, low-resource language focus, and Khmer language research. This screening process led to the elimination of 89 papers deemed insufficiently relevant to their research objectives. Ultimately, 54 papers were identified as most relevant and selected for comprehensive review and analysis to support this study (as illustrated in Figure. A1).

The Studies Analysis and Models Comparison

Following the literature review, a thorough analysis of existing spelling correction models, both traditional and deep learning-based, was conducted. This analysis assessed each model's methodology, strengths, and limitations, with a particular emphasis on their effectiveness in correcting homophone errors, particularly in LRLs. Models were evaluated based on several criteria: context sensitivity to identify those suitable for homophone correction tasks, data dependency to determine applicability to LRLs, performance to assess which models perform best under different conditions, computational requirements to identify models adaptable to resource-constrained environments like LRLs, and bias and data limitations to uncover the weaknesses of each model and

Proceedings DOI: 10.21467/proceedings.8.1
ISBN: 978-81-989164-6-4

Series: AIJR Proceedings
ISSN: 2582-3922

determine their suitability for our needs. The results of this review and analysis provided valuable insights that helped inform the development of a model selection framework specifically designed for resource-constrained settings and the unique challenges of homophone correction.

Case Study: Khmer Spelling Correction

A case study approach was employed to explore the performance of different NLP models in correcting homophone errors in Khmer, drawing insights from existing research. Rather than focusing on experiments or benchmarks, the study emphasizes a conceptual analysis of the challenges in developing such models. This perspective provides a practical and context-sensitive approach to model selection for homophone spelling correction (HSC) systems in low-resource languages (LRLs).

Khmer, the official language of Cambodia, has a complex script consisting of 114

Unicode characters, stacked consonants, and no spaces between words (Buoy et al., 2023; Unicode Consortium, 2025). These features, along with a high frequency of homophones, present significant challenges for NLP, particularly in HSC. Accurate correction requires deep contextual understanding due to frequent word ambiguity. For example, homophone pairs like កា/ការ/ការណ៍/កាណ៌ (ewer/to protect/event/ear - royal word) and words with the suffix ន៍[1], such as អនុវត្ត/អនុវត្តន៍ (to apply/application) and សស/សសន៍ (to interview/interview), are frequently confused by users, creating significant challenges for Khmer speakers (Born et al., 2024).

As an LRL, Khmer faces a shortage of annotated data, making it difficult to train robust models, especially for tasks that require nuance, such as homophone correction. Additionally, computational constraints limit the use of resource-heavy approaches like large language models (LLMs). While simpler models are more efficient, they often lack the context awareness necessary for accurate corrections. With the growing digital presence of Khmer, there is a need for scalable and maintainable models that can adapt to new data, support domain-specific usage, and allow for incremental updates without requiring full retraining. Overcoming these interrelated challenges is crucial for advancing homophone spelling correction in Khmer and other LRLs.

Development Guidelines for Homophone Correction for LRLs

The proposed guidelines for HSC in LRLs are grounded in an extensive review of existing literature, emphasizing crucial factors for effective model development. These guidelines are flexible and suitable for different development stages. They apply to early development with limited resources, cases with moderate data, and situations with ample data and computing power. The proposed guidelines provide a structured approach to developing HSC systems, ensuring that models are tailored to the unique challenges and constraints of LRLs. In conclusion, the methodology applied in this study provides a structured approach to model selection for HSC in LRLs. Through a systematic review of the literature, detailed model analysis, and a case study focused on Khmer, the study lays the foundation for the development of practical guidelines. These guidelines offer actionable insights into selecting and developing models that balance efficiency, performance, and computational

[1] "ន៍" is a suffix that transforms a verb into a noun. For example: "អនុវត្ត" is a verb meaning *to apply*. When the suffix "ន៍" is added, it becomes "អនុវត្តន៍" a noun meaning *application*.

Proceedings DOI: 10.21467/proceedings.8.1
ISBN: 978-81-989164-6-4

Series: AIJR Proceedings
ISSN: 2582-3922

constraints, specifically addressing the challenges posed by homophone errors in resource-constrained settings.

Results

Through a comprehensive analysis of 54 studies selected from an original set of 174, this research uncovered various model capabilities suited to different contexts. It also proposed practical guidelines for developing HSC models in LRLs. The findings serve as a valuable resource for future research, as detailed in the following subsections.

Models Capabilities for Homophone Correction in LRLs

This study analyzed 54 research papers to identify which model capabilities are best suited to various contexts, as summarized in Table B1. The table defines the criteria used to assess different spelling correction models, focusing on factors such as context sensitivity (Con. Sen.), data requirements (Data dep.), performance, computational requirements (Comp. Req.), and bias and data limitations (B&D Lim.). It serves two main functions: first, to synthesize findings from the systematic literature review (SLR) on the strengths and limitations of traditional models (TMs), deep learning models (DLMs), and large language models (LLMs) in correcting homophone errors (HSC); and second, to justify the choice of models for the Khmer case study by aligning specific constraints and opportunities in Khmer with each of these criteria. By organizing the models based on these key factors, Table B1 offers a useful framework for researchers to select the most suitable models for HSC in low-resource languages (LRLs).

Regarding context sensitivity, TMs, like noisy channel and statistical language models, rely on local context and word frequencies, limiting their ability to disambiguate homophones effectively (Eger et al., 2016; Lee et al., 2017). In contrast, DLMs, such as sequence-to-sequence and Transformer-based architectures, effectively capture local and global dependencies, enhancing homophone correction (Büyük & Arslan, 2021; Shah & De Melo, 2020). LLMs, pre-trained on vast multilingual corpora, provide strong context sensitivity through discourse-level understanding, making them highly effective for resolving homophone ambiguities (Allamong et al., 2025; Hu et al., 2020). In terms of data dependency , TMs require minimal annotated data, making them suitable for low-resource settings (Eger et al., 2016; Lee et al., 2017). DLMs require substantial annotated datasets or data augmentation, increasing their data dependency (Büyük & Arslan, 2021; Salhab & Abu-Khzam, 2024). LLMs reduce reliance on task-specific data but still face challenges in LRLs, as their pre-training data is largely from high-resource languages (HRLs), potentially impacting performance (Allamong et al., 2025; Shanahan, 2024). Regarding performance, TMs are computationally efficient but struggle with complex homophone errors due to limited context (Eger et al., 2016; Lee et al., 2017). DLMs improve performance by modeling complex phonology and orthography relationships in homophones (Lee et al., 2020; Shah & De Melo, 2020). LLMs offer the best performance in homophone correction due to their advanced context modeling, although their effectiveness can be reduced by biases from HRL data (Allamong et al., 2025; Tran et al., 2021). In terms of computational requirements , TMs are lightweight and suitable for low-resource environments (Eger et al., 2016; Lee et al., 2017), while DLMs require significant computational power, such as GPUs and extended training times, but provide better results (Büyük & Arslan, 2021; Shah & De Melo, 2020). LLMs are the most computationally expensive, needing

Proceedings DOI: 10.21467/proceedings.8.1
ISBN: 978-81-989164-6-4

Series: AIJR Proceedings
ISSN: 2582-3922

substantial resources for training and fine-tuning, posing challenges in resource-limited settings (Allamong et al., 2025; Shanahan, 2024). Finally, for bias and data limitations, TMs have the advantage of being less biased, yet they struggle with homophone-related errors (Eger et al., 2016; Lee et al., 2017). DLMs can inherit biases from limited or non-representative training data, affecting their ability to correct homophones in LRLs (Salhab & Abu-Khzam, 2024). LLMs reduce data reliance but often exhibit bias from the over-representation of HRLs in their training data, limiting performance in LRLs (Allamong et al., 2025; Shanahan, 2024).

Building on these insights, we apply them to our case study, which produces the following outcomes. TMs performed poorly, especially in contextual understanding. LLMs were more capable but limited by high data and computational demands and biases from HRL-dominated training. DLMs outperformed both, meeting most requirements. Their main drawback, bias from unbalanced data can be mitigated with higher-quality, representative datasets. Overall, DLMs showed the best performance for Khmer homophone correction. In conclusion, each model has unique strengths and limitations in correcting homophone errors in LRLs. Understanding these differences is key to selecting the most suitable model that fits specific needs, resources, and contexts.

This ensures optimal effectiveness in addressing the unique challenges of LRL settings.

The Practical Guidelines for Model Development

Based on the findings presented above, we recommend three key guidelines for developing HSC models in LRLs, presented in Figure A2. First, in resource-constrained environments, TMs like dictionary-based or rule-based systems offer a cost-effective, low-complexity solution, though they have limited context sensitivity (Do et al., 2021; Eger et al., 2016). These models are ideal for early-stage setups due to their minimal data and computational requirements. Second, when moderate amounts of data are available but computing resources remain limited, DLMs such as LSTMs and transformers, enhanced with transfer learning, provide improved accuracy by capturing complex relationships between homophones (Büyük & Arslan, 2021; Etoori et al., 2018). These models also benefit from data augmentation strategies. Third, in scenarios where both large datasets and substantial computing power are accessible, LLMs like GPT and BERT excel in homophone correction due to their strong contextual understanding (Hladek et al., 2020; Jiang et al., 2024). Fine-tuning these models can further improve their performance for homophone correction (He et al., 2023). These guidelines provide a structured approach to developing effective and efficient HSC systems, addressing both technical and resource-related challenges in LRL contexts.

In conclusion, our SLR has uncovered important understandings about the strengths and weaknesses of different model types. Our research emphasizes how crucial it is to choose suitable models according to resource availability and language-specific contexts. These insights and recommendations represent a valuable contribution to the field, offering researchers a more effective framework for model selection, particularly when addressing HSC challenges in LRLs.

Discussion

This review highlights that traditional approaches, such as dictionary- and rule-based models, struggle with homophone errors due to their inability to capture context. Deep learning models (DLMs), like LSTMs and transformers, provide better performance by recognizing complex patterns,

Proceedings DOI: 10.21467/proceedings.8.1
ISBN: 978-81-989164-6-4

Series: AIJR Proceedings
ISSN: 2582-3922

while large language models (LLMs) like GPT and BERT offer superior context awareness. However, LLMs require high computational resources and may introduce biases, especially when applied to low-resource languages (LRLs) like Khmer, where regional dialects and informal speech are underrepresented. Given the data and resource limitations in LRLs, DLMs are currently the most suitable choice for Khmer homophone spelling correction (HSC), balancing accuracy and efficiency. As resources grow, transitioning to LLMs could further enhance performance. The recommended approach is to start with traditional methods and gradually adopt DLMs as resources improve. Bias in NLP models, particularly against non-standard dialects or regional variations, poses a challenge for fairness. Future models must ensure inclusivity by addressing these biases. Additionally, sustainability concerns with LLMs in resource-constrained environments suggest a need for more efficient models. Techniques like transfer learning or model distillation could help make HSC systems more accessible and computationally feasible in LRLs.

Conclusion

Addressing the challenges of homophone spelling correction (HSC) in low-resource languages (LRLs) is critical for improving NLP applications in underrepresented languages such as Khmer. This study highlights the importance of selecting the right language models for these tasks and provides practical guidelines for researchers working in resource-constrained environments. The implications of this work extend beyond Khmer, offering valuable insights for other LRLs facing similar challenges in spelling correction. By providing a structured methodology for model selection and emphasizing the importance of context-aware models, this research paves the way for more accurate and scalable HSC systems in LRLs. Future research should explore the integration of more diverse datasets, especially for languages with limited resources like Khmer, to improve model performance. Additionally, research into computationally efficient methods, such as smaller pre-trained models or transfer learning, could provide significant advantages for real-world applications. Investigating the ethical implications of AI in low-resource settings and developing models that account for regional dialects and sociolects is another promising direction for future work.

References

Allamong, M. B., Jeong, J., & Kellstedt, P. M. (2025). Spelling correction with large language models to reduce measurement error in open-ended survey responses. *Research & Politics*, *12*(1), 20531680241311510. https://doi.org/10.1177/20531680241311510

Attia, M., Pecina, P., Samih, Y., Shaalan, K., & Van Genabith, J. (2016). Arabic spelling error detection and correction. *Natural Language Engineering*, *22*(5), 751–773. https://doi.org/10.1017/S1351324915000030

Born, S., May, M., Piau-Toffolon, C., & Iksal, S. (2024). A survey on importance of homophones spelling correction model for Khmer authors. https://doi.org/10.48550/arXiv.2411.10477

Born, S., Valy, D., & Kong, P. (2022). Encoder-decoder language model for Khmer handwritten text recognition in historical documents. *2022 14th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, 234–238. https://doi.org/10.1109/SKIMA57145.2022.10029532

Buoy, R., Iwamura, M., Srun, S., & Kise, K. (2023). Toward a low-resource non-Latin-complete baseline: An exploration of Khmer optical character recognition. *IEEE Access*, *11*, 128044–128060. https://doi.org/10.1109/ACCESS.2023.3332361

Büyük, O., & Arslan, L. M. (2021). Learning from mistakes: Improving spelling correction performance with automatic generation of realistic misspellings. *Expert Systems*, *38*(5), e12692. https://doi.org/10.1111/exsy.12692

Cissé, T. I., & Sadat, F. (2023). Automatic spell checker and correction for under-represented spoken languages: Case study on Wolof. https://doi.org/10.48550/arXiv.2305.12694

Do, D.-T., Nguyen, H. T., Bui, T. N., & Vo, D. H. (2021). VSEC: Transformer-based model for Vietnamese spelling correction. *PRICAI 2021: Trends in Artificial Intelligence*, 259–272. https://doi.org/10.1007/978-3-030-89363-7_20

Dong, R., Yang, Y., & Jiang, T. (2019). Spelling correction of non-word errors in Uyghur–Chinese machine translation. *Information*, *10*(6), 202. https://doi.org/10.3390/info10060202

Eger, S., vor der Brück, T., & Mehler, A. (2016). A comparison of four character-level string-to-string translation models for (OCR) spelling error correction. *Prague Bulletin of Mathematical Linguistics*. https://doi.org/10.1515/pralin-2016-0004

Eskander, R., Klavans, J., & Muresan, S. (2019). Unsupervised morphological segmentation for low-resource polysynthetic languages. *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 189–195. https://doi.org/10.18653/v1/W19-4222

Etoori, P., Chinnakotla, M., & Mamidi, R. (2018). Automatic spelling correction for resource-scarce languages using deep learning. *Proceedings of ACL 2018, Student Research Workshop*, 146–152. https://doi.org/10.18653/v1/P18-3021

Fahrudin, T. M., Sa'diyah, I., Latipah, L., Atha Illah, I. Z., Bey Lirna, C. C., & Acarya, B. S. (2021). KEBI 1.0: Indonesian spelling error detection system for scientific papers using dictionary lookup and Peter Norvig spelling corrector. *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi*, *12*(2), 78. https://doi.org/10.24843/LKJITI.2021.v12.i02.p02

Goslin, K., & Hofmann, M. (2022). English language spelling correction as an information retrieval task using Wikipedia search statistics. *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, 458–464. https://aclanthology.org/2022.lrec-1.48/

Gueddah, H., & Lachibi, Y. (2023). Arabic spellchecking: A depth-filtered composition metric to achieve fully automatic correction. *International Journal of Electrical and Computer Engineering*, *13*(5), 5366–5373. https://doi.org/10.11591/ijece.v13i5.pp5366-5373

Gueddah, H., Nejja, M., Iazzi, S., Yousfi, A., & Aouragh, S. L. (2022). Improving spellchecking: An effective ad-hoc probabilistic lexical measure for general typos. *Indonesian Journal of Electrical Engineering and Computer Science*, *27*(1), 521–527. https://doi.org/10.11591/ijeecs.v27.i1.pp521-527

Hasan, S., Heger, C., & Mansour, S. (2015). Spelling correction of user search queries through statistical machine translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 451–460. https://doi.org/10.18653/v1/D15-1051

He, Z., Zhu, Y., Wang, L., & Xu, L. (2023). UMRSpell: Unifying the detection and correction parts of pre-trained models towards Chinese missing, redundant, and spelling correction. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10238–10250. https://doi.org/10.18653/v1/2023.acl-long.570

Hladek, D., Staš, J., & Pleva, M. (2020). Survey of automatic spelling correction. *Electronics*, *9*, 1670. https://doi.org/10.3390/electronics9101670

Hu, Y., Jing, X., Ko, Y., & Rayz, J. T. (2020). Misspelling correction with pre-trained contextual language model. *2020 IEEE 19th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC)*, 144–149. https://doi.org/10.1109/ICCICC50026.2020.9450253

Jiang, L., Shen, X., Zhao, Q., & Yao, J. (2024). MLSL-spell: Chinese spelling check based on multi-label annotation. *Applied Sciences*, *14*(6), 2541. https://doi.org/10.3390/app14062541

Kasmaiee, S., Kasmaiee, S., & Homayounpour, M. (2023). Correcting spelling mistakes in Persian texts with rules and deep learning methods. *Scientific Reports*, *13*(1), 19945. https://doi.org/10.1038/s41598-023-47295-2

Kim, J., Weiss, J. C., & Ravikumar, P. (2022). Context-sensitive spelling correction of clinical text via conditional independence. *Conference on Health, Inference, and Learning*, 234–247. https://proceedings.mlr.press/v174/kim22b.html

Kusuma, A. T. A., & Ratnasari, C. I. (2023). Comparison of spell correction in Bahasa Indonesia: Peter Norvig, LSTM, and n-gram. *Jurnal Informatika dan Komputer*, *6*(3), 214–220. https://doi.org/10.33387/jiko.v6i3.7072

Kuznetsov, A., & Urdiales, H. (2021). Spelling correction with denoising transformer. https://doi.org/10.48550/arXiv.2105.05977

Lai, K. H., Topaz, M., Goss, F. R., & Zhou, L. (2015). Automated misspelling detection and correction in clinical free-text records. *Journal of Biomedical Informatics*, *55*, 188–195. https://doi.org/10.1016/j.jbi.2015.04.008

Lee, J.-H., Kim, M., & Kwon, H.-C. (2017). Improved statistical language model for context-sensitive spelling error candidates. *Journal of Korea Multimedia Society*, *20*(2), 371–381.

Lee, J.-H., Kim, M., & Kwon, H.-C. (2020). Deep learning-based context-sensitive spelling typing error correction. *IEEE Access*, *8*, 152565–152578.

Lertpiya, A., Chalothorn, T., & Chuangsuwanich, E. (2020). Thai spelling correction and word normalization on social text using a two-stage pipeline with neural contextual attention. *IEEE Access*, *8*, 133403–133419. https://doi.org/10.1109/ACCESS.2020.3010828

Liang, Z., Quan, X., & Wang, Q. (2023). Disentangled phonetic representation for Chinese spelling correction. https://doi.org/10.48550/arXiv.2305.14783

Lin, Y., Zhang, Z., Hu, M., Sun, Y., & Zhang, Y. (2024). Modalities should be appropriately leveraged: Uncertainty guidance for multimodal Chinese spelling correction. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 11463–11474.

Liu, L., Wu, H., & Zhao, H. (2024). Chinese spelling correction as rephrasing language model. *Proceedings of the AAAI Conference on Artificial Intelligence*. https://doi.org/10.48550/arXiv.2308.08796

Liu, S., Yang, T., Yue, T., Zhang, F., & Wang, D. (2021). PLOME: Pre-training with misspelled knowledge for Chinese spelling correction. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2991–3000. https://doi.org/10.18653/v1/2021.acl-long.233

Liu, Z., Richardson, C., Hatcher, R., & Prud'hommeaux, E. (2022). Not always about you: Prioritizing community needs when developing endangered language technology. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3933–3944. https://doi.org/10.18653/v1/2022.acl-long.272

Magueresse, A., Carles, V., & Heetderks, E. (2020). Low-resource languages: A review of past work and future challenges. https://doi.org/10.48550/arXiv.2006.07264

Mainsah, B. O., Morton, K. D., Collins, L. M., Sellers, E. W., & Throckmorton, C. S. (2015). Moving away from error-related potentials to achieve spelling correction in P300 spellers. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *23*(5), 737–743. https://doi.org/10.1109/TNSRE.2014.2374471

Mammadov, S. (2019). Neural spelling correction for Azerbaijani language. *Proceedings of the 2019 IEEE 13th International Conference on Application of Information and Communication Technologies (AICT)*, 1–5. https://doi.org/10.1109/AICT47866.2019.8981776

Mao, M., Peng, S., Yang, Y., & Park, D.-S. (2022). Bi-directional maximal matching algorithm to segment Khmer words in sentence. *Journal of Information Processing Systems*, *18*(4), 549–561. https://doi.org/10.3745/JIPS.04.0250

Mashod Rana, M., Tipu Sultan, M., Mridha, M. F., Eyaseen Arafat Khan, M., Masud Ahmed, M., & Abdul Hamid, M. (2018). Detection and correction of real-word errors in Bangla language. *Proceedings of the 2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, 1–4. https://doi.org/10.1109/ICBSLP.2018.8554502

Moslem, Y., Haque, R., & Way, A. (2020). Arabisc: Context-sensitive neural spelling checker. *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications*, 11–19. https://doi.org/10.18653/v1/2020.nlptea-1.2

Mukazhanov, N., Alibiyeva, Z., Yerimbetova, A., Kassymova, A., & Alibiyeva, N. (2023). Development of an augmented Damerau–Levenshtein method for correcting spelling errors in Kazakh texts. *Eastern-European Journal of Enterprise Technologies*, *5*(2), 23–33. https://doi.org/10.15587/1729-4061.2023.289187

Phung, H. T., & Luong, N. V. (2024). Detecting spelling errors in Vietnamese administrative document using large language models. *Ho Chi Minh City Open University Journal of Science – Engineering and Technology*, *14*(1), 31–40. https://doi.org/10.46223/HCMCOUJS.tech.en.14.1.3141.2024

Salhab, M., & Abu-Khzam, F. (2024). AraSpell: A deep learning approach for Arabic spelling correction. https://doi.org/10.48550/arXiv.2405.06981

San, M., Cai, Z., Cai, R., & Dao, J. (2021). Analysis on types of spelling errors in true Tibetan characters. *MATEC Web of Conferences*, *336*, 06019. https://doi.org/10.1051/matecconf/202133606019

Shah, K., & De Melo, G. (2020). Correcting the autocorrect: Context-aware typographical error correction via training data augmentation. https://doi.org/10.48550/arXiv.2005.01158

Shanahan, M. (2024). Talking about large language models. *Communications of the ACM*, *67*(2), 68–79. https://doi.org/10.1145/3624724

Sun, R., Wu, X., & Wu, Y. (2023). An error-guided correction model for Chinese spelling error correction. https://doi.org/10.48550/arXiv.2301.06323

Sung, T., & Hwang, I. (2016). Ternary decomposition and dictionary extension for Khmer word segmentation. *Journal of Information Technology Applications and Management*, *23*(2), 11–28. https://doi.org/10.21219/JITAM.2016.23.2.011

Tran, H., Dinh, C. V., Phan, L., & Nguyen, S. T. (2021). Hierarchical transformer encoders for Vietnamese spelling correction. *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, 547–556. https://doi.org/10.48550/arXiv.2105.13578

Unicode Consortium. (2025). *The Unicode standard, version 16.0*. Retrieved July 14, 2025, from https://www.unicode.org/charts/PDF/U1780.pdf

Van Nam, T., Thi Hue, N., & Huy Khanh, P. (2017). Building a syllable database to solve the problem of Khmer word segmentation. *International Journal on Natural Language Computing*, *6*(1), 1–12. https://doi.org/10.5121/ijnlc.2017.6101

Zhang, R., Pang, C., Zhang, C., Wang, S., He, Z., Sun, Y., Wu, H., & Wang, H. (2021). Correcting Chinese spelling errors with phonetic pre-training. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2250–2261. https://doi.org/10.18653/v1/2021.findings-acl.198

Zhang, S., Huang, H., Liu, J., & Li, H. (2020). Spelling error correction with soft-masked BERT. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 882–890. https://doi.org/10.18653/v1/2020.acl-main.82

Zhou, Y., Porwal, U., & Konow, R. (2017). Spelling correction as a foreign language. https://doi.org/10.48550/arXiv.1705.07371

Appendix A

Figures

Figure A1

The Paper Selection Process

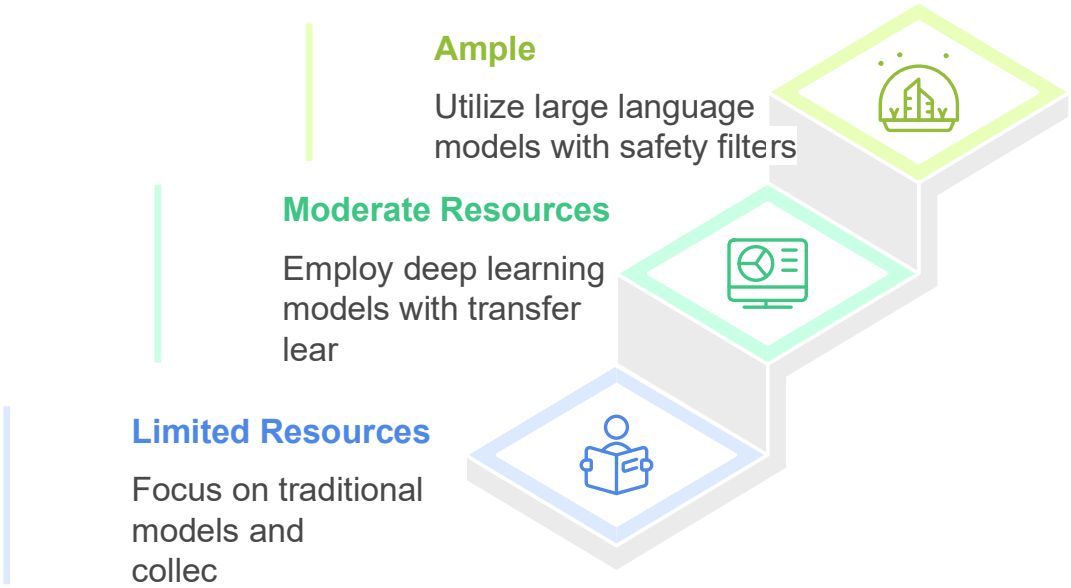

Figure A2

Resource-aware Guideline for HSC in LRLs.

Proceedings DOI: 10.21467/proceedings.8.1
ISBN: 978-81-989164-6-4

Series: AIJR Proceedings
ISSN: 2582-3922

Appendix B

Tables

Table B1

Comparison of models for homophone spelling correction in low-resource languages

Criteria	TMs	DLMs	LLMs
Con. Sen.	Contextual ambiguity (Eger & Mehler, 2016; Lee et al., 2017)	Context-aware modeling (Buyuk et al., 2021; Shah & Lee, 2020)	Discourse-level understanding (Allamong et al., 2025; Hu et al., 2021)
Data Dep.	Data-efficient modeling (Eger & Mehler, 2016; Lee et al., 2017)	Data-intensive modeling (Buyuk et al., 2021; Salhab et al., 2023)	Pretrained model bias (Allamong et al., 2025; Shanahan, 2024)
Performance	Contextual limitation (Eger & Mehler, 2016; Lee et al., 2017)	Phonological awareness (Shah & Lee, 2020; Lee et al., 2020)	Advanced contextual modeling (Allamong et al., 2025; Tran et al., 2021)
Comp. Req.	Lightweight modeling (Eger & Mehler, 2016; Lee et al., 2017)	Computationally intensive (Buyuk et al., 2021; Shah & Lee, 2020)	Resource-intensive modeling (Allamong et al., 2025; Shanahan, 2024)
B&D Lim.	Bias-resistant but context-limited (Eger & Mehler, 2016; Lee et al., 2017)	Data-induced bias (Salhab et al., 2023)	Cross-lingual transfer bias (Allamong et al., 2025; Shanahan, 2024)

Note. This table compares the strengths and limitations of three model categories in homophone spelling correction for LRLs, focusing on contextual handling, data needs, model performance, computational demand, and bias susceptibility.

Proceedings DOI: 10.21467/proceedings.8.1
ISBN: 978-81-989164-6-4

Series: AIJR Proceedings
ISSN: 2582-3922