Morphology-Aware Retrieval for Low-Resource Environments: Advancing Information Retrieval for Shona Language

Ruvimbo Maud Munetsi
Department of Computer Science
University of Zimbabwe [1]
Harare Institute of Technology [2]
Harare, Zimbabwe
rmunetsi@hit.ac.zw

Tendai Mukande
Insight Research Ireland Centre for Data Analytics
Dublin City University
Dublin, Ireland
tendai.mukande2@mail.dcu.ie

Noel O'Connor
Insight Research Ireland Centre for Data Analytics
Dublin City University
Dublin, Ireland
Noel.OConnor@dcu.ie

Abstract

Research on Information Retrieval (IR) has historically prioritised high-resource languages such as English and Chinese, with less attention given to many low-resource languages. For example, Shona, a Bantu language spoken by approximately 12 million people in Zimbabwe and neighbouring countries, remains under-explored in IR research despite its widespread societal use in Southern Africa. In this work, we present a preliminary study of Shona IR using sparse and dense retrieval models, demonstrating significant performance limitations due to morphological complexity and data scarcity. Based on these findings, we propose to develop a framework to advance Shona IR by developing a large-scale benchmark dataset to support morphology-aware retrieval. We hypothesise that improving Shona IR supports equitable access to digital information and enables language-inclusive AI technologies aligned with global development priorities such as accessibility to education, dissemination of healthcare information, and digital inclusion.

CCS Concepts

• **Information systems** → **Specialized information retrieval**; **Retrieval models and ranking**; **Retrieval tasks and goals**.

Keywords

Information Retrieval, Low-Resource Languages, Shona, Morphology-Aware Retrieval, Benchmark Datasets

ACM Reference Format:

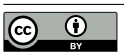
Ruvimbo Maud Munetsi, Tendai Mukande, and Noel O'Connor. 2026. Morphology-Aware Retrieval for Low-Resource Environments: Advancing Information Retrieval for Shona Language. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3805712.3808522

1 Introduction

Effective Information Retrieval (IR) plays a vital role in promoting equitable access to digital knowledge and services. However, most modern IR models have been developed primarily for high-resource languages, leaving many low-resource indigenous languages significantly under-represented [3]. Low-resource languages pose distinct challenges for modern IR models [16]. A primary limitation is the scarcity of large-scale annotated corpora and standardised evaluation benchmarks [27]. Moreover, the available digital resources for these languages are largely concentrated in formal domains, such as religious texts, educational materials, and curated news content [20]. These datasets often fail to capture informal writing styles and domain-specific terminology encountered in real-world IR scenarios [10]. Furthermore, their scarcity restricts robust evaluation and benchmarking, thus limiting progress in academic research and industry [2]. In this work, we focus on the Shona language, spoken by more than 12 million people in Southern Africa. As in other Bantu languages, Shona exhibits rich morphological variation by using noun class markers, prefixes, suffixes, and verb extensions [19, 22]. Consequently, a single root word can surface in multiple forms depending on grammatical tense or semantic modification [26].

We define morphology-aware retrieval as the ability of a model to retrieve semantically relevant documents despite the variation in surface forms caused by inflection, derivation, or agreement. The benchmark will therefore evaluate models on their ability to generalise across such variations. Traditional lexical retrieval approaches, such as BM25 [25], rely on exact or near-exact token matching and therefore struggle to capture semantic equivalence between related terms [9, 24]. For example, the Shona word *munhu*(person) transforms into *vanhu* (people) through noun-class prefix substitution rather than by simple suffix addition. Although both forms share the root "-nhu", models often treat them as distinct tokens, unless morphology-aware representations are applied [8]. In addition, real-world Shona communication often involves code-switching with English and other regional languages, further complicating query understanding and document retrieval [12]. Consequently, most existing IR models struggle to generalise in diverse linguistic contexts, limiting their practical deployment in applications such as education, healthcare, and public service [6].

Our preliminary analysis of a collection of online Shona documents reveals several challenges that motivate this work. Standard lexical retrieval models such as BM25 struggle to match different inflected forms of the same root, leading to the omission of relevant documents. For example, a query containing the verb *"kudzidza"* ("to learn") fails to retrieve documents containing its inflected form *"adzidzira"* ("has learned"), even though the content is semantically relevant. In addition, queries that include English terms, such as *"Shona mathematics textbook"*, are poorly handled, reducing both coverage and relevance. Dense multilingual models such as LLMs, pre-trained on high-resource languages, capture semantic similarity



SIGIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3808522

Table 1: Preliminary Evaluation Results of Selected IR Models on the Shona IR Dataset

Model	Recall@5	Recall@10	nDCG@5	nDCG@10	MAP	MRR	Handling Inflection
BM25 [25]	0.52	0.70	0.45	0.47	0.35	0.52	Low
ColBERT-v2 [28]	0.56	0.73	0.51	0.54	0.40	0.59	Moderate
Voyager [30]	0.60	0.78	0.55	0.58	0.44	0.63	High
Bloom [31]	0.58	0.76	0.54	0.57	0.42	0.61	High
BGE-M3 [6]	0.57	0.75	0.53	0.56	0.41	0.60	Moderate

only partially [15], as Shona-specific training data is scarce, leading to suboptimal retrieval performance. These findings indicate that existing retrieval methods are insufficient for low-resource, morphologically rich languages and highlight the need to address these challenges.

Our goal is to develop a large-scale, publicly accessible Shona IR benchmark dataset that effectively handles morphological variation. The dataset will be curated from various sources, including news articles, government publications, educational materials, and cultural texts. The proposed benchmark aims to address the morphological challenges identified in our preliminary findings to promote equitable access to information and digital inclusion.

2 Background: *Shona* Language

Shona (also called *chiShona*) is a Bantu language of the Niger–Congo family, spoken mainly in Zimbabwe and parts of Mozambique, Botswana and Zambia.[1] Shona draws on central dialects such as Zezuru and Karanga. It is one of Zimbabwe's official languages and is widely used in education, media, and literature. Shona uses a Latin-based alphabet and features a tonal system with high and low tones that distinguish meaning [18, 19]. The language exhibits a complex noun class system, characteristic of Bantu languages, in which nouns are organised into classes marked by prefixes that determine the agreement with verbs and modifiers [22]. The basic word order is *Subject–Verb–Object* (SVO).[2] Shona has five vowel phonemes and a rich consonant inventory including distinctive fricatives such as *sv* and *zv*. In the twentieth century, a standardised orthography was developed, and the language has a strong tradition of oral literature and written works [5].

2.1 Related Work

Previous work has shown that morphological processing techniques, such as segmentation, lemmatisation, and subword modelling, can mitigate sparsity and improve downstream tasks in low-resource settings [7, 8]. For example, transfer-based approaches have been used to learn morphological patterns by leveraging related languages with richer resources [21]. This is relevant for Bantu languages, where shared grammatical structures (e.g., noun class systems and agglutinative morphology) suggest that resources developed for languages such as *isiZulu* or *Sepedi* may be reusable for the Shona language [19].

Existing efforts have begun to develop datasets and benchmarks for African languages, including Amharic [4, 17] and Swahili [11, 23]. However, for many African languages, including Shona, available digital content is largely restricted to formal domains such as religious texts, news articles, and educational materials, with limited representation of conversational or user-generated content. This imbalance reduces the ecological validity of existing datasets and highlights the need for domain-diverse IR benchmarks [16].

3 Methodology

Our preliminary evaluation was conducted on a collection of approximately 5,000 Shona-language documents gathered from publicly accessible sources, including news articles, government publications, educational materials, and cultural texts. The dataset is designed to reflect diverse domains and realistic language use, including both formal and semi-formal content. We construct a set of 75 information needs (topics) to simulate realistic user queries. These queries span multiple domains, such as education, healthcare and general knowledge, and include both purely Shona queries and code-switched Shona-English queries. Query formulation is informed by common information-seeking patterns observed in low-resource language settings, where users often employ short, underspecified queries and frequently engage in code-switching due to limited standardised terminology [1, 20]. To reflect these characteristics, the query set includes both monolingual Shona queries and mixed Shona–English expressions in domains such as education, healthcare, and general knowledge.

Relevance judgments are obtained through manual annotation by fluent Shona speakers to ensure linguistic and contextual accuracy. We use human annotators as automatic labelling methods may fail to capture semantic equivalence in inflected or derived forms [26]. Annotators are instructed to assess relevance based on topical alignment and the degree to which a document satisfies the underlying information need, rather than relying on exact lexical overlap. For each query, the top 10 documents are merged to form a candidate pool, which is then manually annotated. This approach provides a cost-effective mechanism to build relevance judgments while maintaining adequate coverage of high-ranking results [13, 29].

3.1 Evaluation Protocol

We evaluate retrieval performance using Normalised Discounted Cumulative Gain (nDCG), Mean Average Precision (MAP), and Mean Reciprocal Rank (MRR). We apply a range of retrieval approaches, including: (i) sparse lexical retrieval (BM25), (ii) late-interaction neural retrieval (ColBERT-v2), and (iii) dense semantic retrieval models (Voyager, BLOOM, and BGE-M3). To reflect the

[1] Wikipedia: Shona language.
[2] Shona Language — Dialects & Structure (MustGo).

morphological richness of Shona, the evaluation includes targeted query-document pairs exhibiting:

- **Inflectional variation**, where queries and relevant documents contain different surface forms of the same root (e.g., verb tense or agreement markers);
- **Noun class variation**, including singular–plural transformations (e.g., *munhu* vs. *vanhu*);
- **Derivational morphology**, where related meanings are expressed through affixation;
- **Code-switching**, where queries mix Shona and English terms.

We hypothesise that this approach enables systematic evaluation of whether models can capture semantic equivalence across morphologically related forms, rather than relying solely on surface-level token matching. All models are used in their pre-trained form without fine-tuning on Shona-specific data. We do not perform additional fine-tuning in this preliminary study, allowing us to assess out-of-the-box performance in a low-resource setting.

4 Initial Findings

From preliminary evaluation, we observed a scarcity of annotated text collections and standard benchmarks, which limits the reliability of model comparison. We attribute this to several factors, including limited linguistic resources (e.g. dictionaries and grammars), the scarcity of digitally available textual data, and the underdeveloped infrastructure for large-scale annotation [15]. In addition, a range of dialects increases the difficulty of preprocessing (tokenisation, lemmatisation) and evaluation when data is limited [14].

Evaluation Results. Table 1 presents the performance of 5 models. We use metrics such as Recall, Normalised Discounted Cumulative Gain (NDCG), Mean Average Precision (MAP), and Mean Reciprocal Rank (MRR) to evaluate ranking quality. BM25 produces lower MAP and nDCG scores, highlighting its inability to handle inflectional variation and lexical sparsity effectively. ColBERT-v2 [28] shows notable improvement, while dense retrieval models such as Voyager [30], BLOOM, and BGE-M3 [6] achieve only moderate performance, suggesting that semantic similarity alone is insufficient to address the challenges posed by morphological variation. These results highlight the need for richer datasets to support robust retrieval, motivating the development of approaches that address linguistic complexity and data scarcity.

Morphology Sensitivity Analysis. To assess the impact of morphological variation on retrieval performance, we identify a subset of queries in which relevant documents contain inflected or derived forms of the query terms (e.g., verb extensions and noun class variations). We observe that sparse lexical models such as BM25 fail to retrieve relevant documents when surface forms differ, resulting in lower *Recall*. Dense retrieval models show partial robustness by capturing semantic similarity but degraded performance in cases involving complex morphological transformations. From these findings, we observe that morphological variation is a key source of retrieval error in Shona and motivate the need for morphology-aware evaluation and modelling. These findings suggest that current multilingual models rely largely on cross-lingual transfer and subword representations, yet remain limited in their ability to capture Shona-specific linguistic phenomena effectively.

5 Future Work

We propose a Shona IR benchmark that aims to support practical applications for Shona-speaking communities. For example, in the health domain, queries such as *"Nzira dzekudzivirira malaria kumusha"* ("Ways to prevent malaria in rural areas") allow users to access relevant government advisories, while in education, queries such as *"as "Masvomhu emugwaro rechinomwe kuShona"* ("Grade 7 Mathematics in Shona") allow effective retrieval of curriculum-aligned textbooks and tutorials. We aim to provide a reproducible framework for evaluating retrieval approaches that address morphological challenges, while also supporting real-world information access for Shona-speaking communities.

Morphology-Aware Framework. The proposed framework incorporates morphology-aware evaluation by constructing targeted query–document pairs that systematically reflect controlled linguistic variation. Specifically, the dataset will include: (i) queries whose relevant documents contain different inflectional forms of the same root, (ii) noun class transformations reflecting agreement patterns, (iii) derivational variants formed through affixation, and (iv) code-switched queries combining Shona and English. In addition, the benchmark will provide optional morphological normalisation (e.g., stemming or lemmatisation).

Significance of the Study. The proposed research aligns with several United Nations Sustainable Development Goals (SDGs). Improved Shona IR models would improve access to educational materials and digital learning resources, contributing to *SDG 4*[3]. Enhanced access to health information and public awareness resources supports *SDG 3*[4]. Expanding equitable access to digital services and online information for underrepresented language communities contributes to *SDG 10*[5].

6 Conclusion

In this work, we highlight the challenges posed by Shona's rich morphological structure, where the extensive use of prefixes, suffixes, and noun-class markers often leads to retrieval errors in IR models. We show that the scarcity of high-quality Shona datasets degrades model performance, exacerbating algorithmic bias and digital inequality. To address these issues, we propose a framework to enhance the accuracy and accessibility of digital knowledge for Shona-speaking communities, to bridge the global digital language divide.

Acknowledgments

This research is supported by Taighde Éireann – Research Ireland under Grant Number SFI/12/RC/2289 P2 (Insight Research Ireland Centre for Data Analytics), co-funded by the European Regional Development Fund.

[3]https://sdgs.un.org/goals/goal4
[4]https://sdgs.un.org/goals/goal3
[5]https://sdgs.un.org/goals/goal10

References

[1] Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric NLP for African languages: Where we are and where we can go. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 3814–3841.

[2] David Ifeoluwa Adelani et al. 2022. AfriBERTa: A Multilingual Pretrained Language Model for African Languages. *ACL Findings* (2022).

[3] Akari Asai et al. 2021. Cross-Lingual Open-Retrieval Question Answering. In *ACL*.

[4] Israel Abebe Azime, Mitiku Yohannes Fuge, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Aman Kassahun Wassie, Eyasu Shiferaw Jada, Yonas Chanie, Walelign Tewabe Sewunetie, and Seid Muhie Yimam. 2024. Enhancing amharic-llama: Integrating Task-Specific and Generative Datasets. In *5th Workshop on African Natural Language Processing*.

[5] Michael FC Bourdillon. 1976. *The Shona peoples: An Ethnography of the Contemporary Shona, with Special Reference to Their Religion*. Vol. 1. Mambo press.

[6] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. *arXiv preprint arXiv:2402.03216* 4, 5 (2024).

[7] Santwana Chimalamarri, Dinkar Sitaram, and Ashritha Jain. 2020. Morphological Segmentation to Improve Cross-Lingual Word Embeddings for Low-Resource Languages. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 19, 5 (2020), 1–15.

[8] Eve V Clark. 2017. Morphology in Language Acquisition. *The Handbook of Morphology* (2017), 374–389.

[9] Stéphane Clinchant and Eric Gaussier. 2013. Information Retrieval in Agglutinative Languages. In *CLEF*.

[10] Alexis Conneau et al. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *ACL*.

[11] Samuel Gyamfi, Alfred Malengo Kondoro, Yankı Öztürk, Richard Hans Schreiber, and Vadim Borisov. 2026. Synthetic Data Generation Pipeline for Low-Resource Swahili Sentiment Analysis: Multi-LLM Judging with Human Validation. In *Proceedings of the 7th Workshop on African Natural Language Processing (AfricaNLP 2026)*. 116–141.

[12] Edward Hu et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. *ICLR* (2022).

[13] Minnu Helen Joseph and Sri Devi Ravana. 2024. Reliable Information Retrieval Systems Performance Evaluation: A Review. *IEEE Access* 12 (2024), 51740–51751.

[14] Julia Kreutzer et al. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *TACL* (2022).

[15] Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *NeurIPS*.

[16] Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of Past Work and Future Challenges. *arXiv preprint arXiv:2006.07264* (2020).

[17] Kidist Amde Mekonnen, Yosef Worku Alemneh, and Maarten de Rijke. 2025. Optimized Text Embedding Models and Benchmarks for Amharic Passage Retrieval. In *Findings of the Association for Computational Linguistics: ACL 2025*. 10428–10445.

[18] Lucio Melazzo et al. 2005. Latin Object and Subject Infinitive Clauses. *Universal Grammar in the Reconstruction of Ancient Languages* (2005), 339–372.

[19] Scott P Myers. 1987. *Tone and the structure of words in Shona*. University of Massachusetts Amherst.

[20] Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, et al. 2020. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. In *Findings of EMNLP*.

[21] Toan Q Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 296–301.

[22] Derek Nurse and Gérard Philippson. 2006. *The bantu languages*. Vol. 4. Routledge.

[23] Millicent Ochieng, Varun Gumma, Sunayana Sitaram, Jindong Wang, Vishrav Chaudhary, Keshet Ronen, Kalika Bali, and Jacki O'Neill. 2025. Beyond Metrics: Evaluating LLMs Effectiveness in Culturally Nuanced, Low-Resource Real-World Scenarios. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*. 230–247.

[24] Ari Pirkola. 2001. Morphological Typology of Languages for IR. *Journal of Documentation* 57, 3 (2001), 330–348.

[25] Stephen Robertson, Hugo Zaragoza, et al. 2009. The Probabilistic Relevance Framework: BM25 and beyond. *Foundations and Trends® in information retrieval* 3, 4 (2009), 333–389.

[26] Sebastian Ruder et al. 2023. A Survey of Cross-Lingual Transfer Learning. *JAIR* (2023).

[27] Abhi Ram Reddy Salammagari and Gaurava Srivastava. 2024. Advancing Natural Language Understanding for Low-Resource Languages: Current Progress, Applications, and Challenges. *International Journal of Advanced Research in Engineering and Technology* 15 (2024), 244–255.

[28] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. Colbertv2: Effective and Efficient Retrieval via Lightweight Late Interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 3715–3734.

[29] Ellen Voorhees and Donna Harman. 1998. Overview of the Sixth Text Retrieval Conference (TREC-6). *Nist Special Publication Sp* (1998), 1–24.

[30] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An Open-ended Embodied agent with Large Language Models. *arXiv preprint arXiv:2305.16291* (2023).

[31] BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100* (2022).