Morphologically-informed Somali Lemmatization Corpus built with a Web-based Crowdsourcing Platform

Abdifatah A. Gedi[1] **Shafie A. Mohamed**[1] **Yusuf A. Yusuf**[1]
Muhidin A. Mohamed[1,2] **Fuad M. Hassan**[3] **Houssein A. Assoweh**[4]

[1]Jamhuriya University of Science & Technology, Mogadishu, Somalia; [2]Aston University, UK
[3]Somali National University, Mogadishu, Somalia; [4]Université de Djibouti, Djibouti

Correspondence: Gedi@just.edu.so

Abstract

Lemmatization, which reduces words to their root forms, plays a key role in tasks such as information retrieval, text indexing, and machine learning-based language models. However, a key research challenge for low-resourced languages such as the Somali is the lack of human-annotated lemmatization datasets and reliable ground truth to underpin accurate morphological analysis and training relevant NLP models. To address this problem, we developed the first large-scale, purpose-built Somali lemmatization lexicon, coupled with a crowdsourcing platform for ongoing expansion. The system leverages Somali's agglutinative and derivational morphology, encompassing over 5,584 root words and 78,629 derivative forms, each annotated with part-of-speech tags. For data validation purpose, we have devised a pilot lexicon-based lemmatizer integrated with rule-based logic to handle out-of-vocabulary terms. Evaluation on a 294-document corpus covering news articles, social media posts, and short messages shows lemmatization accuracies of 51.27% for full articles, 44.14% for excerpts, and 59.51% for short texts such as tweets. These results demonstrate that combining lexical resources, POS tagging, and rule-based strategies provides a robust and scalable framework for addressing morphological complexity in Somali and other low-resource languages.

1 Introduction

Lemmatization is a foundational step in Natural Language Processing (NLP) which supports tasks such as information retrieval, text classification, and machine translation by reducing words to their canonical forms. It is a morphological process that converts inflected words to their base forms, also known as lemmas. For nouns, this corresponds to the singular form; for verbs, the infinitive; and for adjectives or adverbs, the positive form. Essentially, lemmatization normalizes different morphological variants of a word by mapping them to the same underlying lemma, allowing them to be analyzed as a single term or concept. By reducing the number of distinct terms, lemmatization simplifies text and benefits downstream processing tasks. For example, in information retrieval systems, lemmatization can improve recall, as queries and documents that are morphologically normalized are more likely to match (Liu et al., 2012).

For many low-resource African languages, lemmatization is particularly challenging due to sparse annotated data and intricate morphological patterns (Adelani, 2025). Recent research has made progress through approaches such as multilingual pre-training, morphological segmentation, and limited supervised datasets, but overall performance remains inconsistent. This paper examines the state of the lemmatization task and resource for the Somali language.

Prior work on Somali lemmatization has been limited, characterized by minimal data and no integration of part-of-speech (POS) information, despite its importance for handling Somali's complex morphology (Mohamed and Mohamed, 2023). In this paper, we expand the dataset, introduce an annotation tool to streamline labeling, and propose a hybrid approach combining rule-based methods, root derivation, lookup strategies, and POS tagging. Our evaluation demonstrates significant improvements over earlier methods, establishing a more robust foundation for future Somali NLP research in low-resource conditions.

In this study, we address the problem of lemmatization for the Somali language, aiming to develop a method for normalizing words derived from the same root. Our focus is primarily on the "MAXAA TIRI," – the principal written dialect of Somali – which was previously explored in our previous initial study (Mohamed and Mohamed, 2023). Building on that work, we have signifcantly expanded the datasets in terms of the root words, their derivative

forms, and the addition of a purposeful annotation tool to facilitate the creation of high-quality linguistic resources supporting crowdsourcing and future related NLP research. The pilot lemmatizer built on the developed dataset identifies and extracts meaningful root forms from inflected variants, employing a hybrid approach that integrates lookup method with rule-based processing and providing a robust foundation for further computational processing of Somali word normalization.

The main contributions of this work are as follows:

1. First, we constructed an expanded Somali lemmatization lexicon that integrates Somali morphological rules and covers a wider range of root forms and their inflections.

2. Second, we designed and implemented an annotation tool to enable effective collaboration across the annotation team.

3. Third, we created a Somali word lemmatization algorithm built on the expanded lexicon incorporated with rule-based method

4. Fourth, we tested the lemmatizer on a Somali corpus of various lengths and domains to evaluate its performance.

2 Related Work

Text lemmatization is a fundamental NLP task, which is considered a solved research problem for high-resource languages such as English, French, and Chinese (Bergmanis and Goldwater, 2018; Manjavacas et al., 2019). However, it remains a significant research challenge for under-resourced languages like the Somali (Miletić and Siewert, 2023; Mohamed and Mohamed, 2023). Although significant progress has been made for various NLP tasks with the emergence of neural networks and transformer-based models, the development of manually annotated lemmatization datasets, such as root-derived word pairs enriched with part-of-speech tags, remains a well-established and indispensable approach for languages with complex morphology and limited digital resources, such as Somali (Sahala et al., 2023; Stanković et al., 2016; Gordin et al., 2025).

Linked with the above, several related studies have contributed to the resource development of core text normalization NLP steps including word lemmatization. For example, the recent work of Mathayo and Kondoro (2024) on Swahili, which is a low-resourced language, introduced a large verb conjugation dataset to address its agglutinative morphology, Covering over 319,000 verb forms, this dataset facilitates essential NLP pipeline steps including lemmatization, and morphological analysis, making it a valuable resource for advancing NLP in low-resource Bantu languages. Moreover, KinyaBERT (Nzeyimana and Rubungo, 2022) demonstrates that integrating explicit morphological structure into transformer models improves performance over subword-only methods for Kinyarwanda While multilingual neural parsing studies on Bambara, Wolof, and Yoruba (Dione, 2021) show that neural transfer approaches still rely on structured annotations such as lemmas and morphological features.

Despite the growing interest in NLP for low-resource languages, Somali remains notably underrepresented in the literature. Few studies have addressed core NLP tasks or the development of language resources for the language. For instance, Mohammed (2020) investigated part-of-speech (POS) tagging using statistical and machine learning methods, achieving an accuracy of 87.51% through ten-fold cross-validation. Additionally, (Badel et al., 2023) develop an annotated corpus – a dataset consisting of 2,335 documents sourced from prominent online platforms, including Hiiraan Online, Dhacdo.net, and collections of Somali poetry – for Somali language information retrieval. Also, Nimaan et al. (2006) explored automatic speech transcription for Somali language, constructing a 10-hour audio corpus and reporting a word error rate (WER) of approximately 21%. Recently, Mohamed et al. (2025) developed two Somali datasets for fake news detection and toxicity classification sourced from the social media and labelled by human annotators. Their work has also introduced the first monolingual BERT-based Somali language model, named SomBERTa, which outperformed compared multilingual models like AfriBERTa and AfroXLMR in fake news and toxicity classification, achieving the highest average accuracy of 87.99% and highlighting promising directions for Somali NLP research. Other research studies on low-resourced languages have utilized multilingual LLMs covering Somali NLP tasks such as machine translation (Wang et al., 2024; Adelani et al., 2022) and text classification (Adelani et al., 2023; Alabi et al., 2022).

The current study follows our previous work

(Mohamed and Mohamed, 2023) which pioneered the development of a Somali lemmatization resource and has specifically addressed lemmatization for the Somali language. While this research builds on that initial study (which was released only as a non-archived preprint), it makes several significant extensions, including the expansion of the dataset from 1247 root words to 5584 words, the addition of morphological information such as part-of-speech (POS) tags, and most importantly, the development of a tailored web-based data annotation tool to facilitate community crowdsourcing and further corpus development.

3 Methodology

This section outlines the methodology employed in developing the data management platform, and the comprehensive Somali lexical database for word lemmatization paired with associated PoS tags. The research realization pipeline includes six key phases: system development, data collection, annotation, lexical database construction, crowdsourced validation, and pilot testing the constructed data in the form of lemmatization evaluation.

3.1 Lemmatization platform and database

A custom web-based annotation platform was developed to facilitate the data collection, annotation, and expansion of the dataset beyond this research via community crowdsourcing [1]. Technically speaking, and drawing inspiration from similar platforms and lexical resources (Habash and Dorr, 2003), we implemented a relational database architecture in which each root word serves as a lexical anchor linked to its derived forms and their PoS tags. All data were stored using a structured SQL schema, allowing for efficient retrieval, expansion, and linkage between morphological variants. This system also ensures data quality by enforcing structural consistency, preventing duplicate entries, and standardizing POS categories while allowing annotators to systematically record derivational variants alongside their PoS tags, thereby producing reliable and high quality lexical resource (Figure 1).

This collaborative lemmatization system formalizes the division of tasks among contributors, combining linguistic expertise with scalable annotation practices. By embedding validation and verification at multiple points in the pipeline, the resulting resource achieves a balance between morphological precision and usability for downstream NLP tasks such as lemmatization, POS tagging, and machine translation.

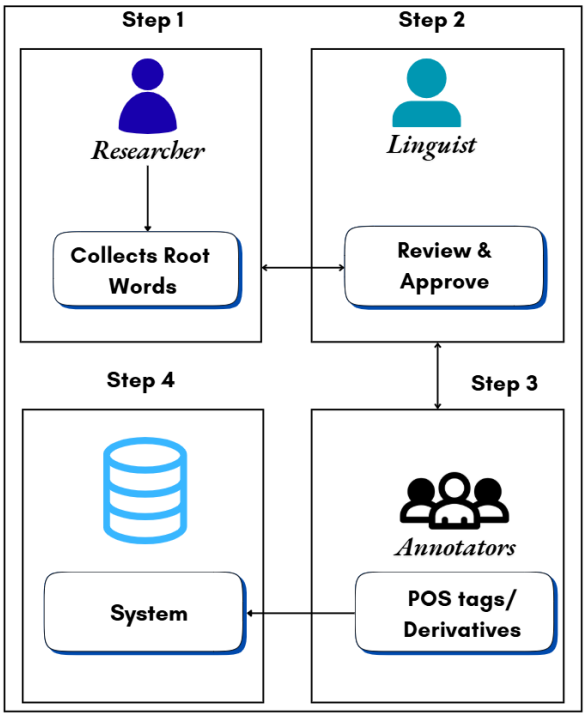

Figure 1: Data annotation and expert approval process

All collected and annotated data, including the compilation of root words, the derived forms, and their PoS tags were iteratively verified, with Somali language experts from the *Intergovernmental Academy for Somali Language* who conducted cross-checks of the data quality, derivational logic, the POS assignments, and overall data consistency. Discrepancies were occasionally identified during these verifications and resolved, which strengthened both the reliability and linguistic validity of the data set.

3.1.1 Core platform features

The developed platform consists of all necessary features from uploading word base forms and annotating them with their derived forms and PoS tags, to data quality moderation and statistical monitoring through tailed dashboard. The root word uploading feature (Figure 7 *appendix A*) serves as a central component of the system and database, enabling insertion, modification, and deletion of root words. Integrated validation mechanisms prevent duplication and enforce conformity with Somali morphological rules. Once approved, root words form the foundation for derivative generation and POS annotation, ensuring accuracy and scalability

[1] http://annot.just.edu.so/

in the lexical resource.

The annotators then start enriching approved root words with derivative forms and assign the corresponding part-of-speech (POS) tag. As shown in (Figure 8 *appendix B*), the interface displays only validated roots, ensuring quality control. Users specify the word type, POS category, root word, and all related derivatives. This structured workflow maintains annotation consistency, prevents duplicates, and links each derivative to its validated root, creating linguistically robust lexical entries that support downstream NLP tasks such as lemmatization and morphological analysis.

The platform also includes a publicly accessible search interface and supports querying root and derived forms, filtering by POS tags, and exporting data for external analysis. For example, it enables users to explore lexical data by entering full or partial terms. The system retrieves root forms, POS tags, and derived forms, including their morphological structures. For example, a search query such as "abuu" returns all words that contain this substring, whether as a root or derived word (Figure 9). This transparent and intuitive interface facilitates open access to the lexical database, enabling linguists, NLP researchers, and language learners to analyze and explore Somali morphology effectively. The system architecture was designed for scalability, allowing integration with additional NLP tools in the future.

Finally, the system is designed to support community crowdsourcing to future expansion of the developed lemmatization resource, enhance scalability and maintain data quality (Figure 2). This would require new users and data annotators to register and undergo approval before contributing to the data expansion. And as with current data, all future submissions need to be reviewed by expert moderators through a cross-verification process to ensure consistency in POS tagging and derivation logic.

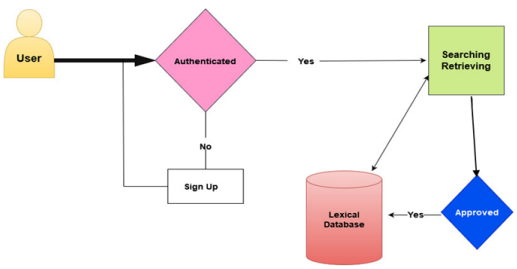

Figure 2: Platform's crowd-sourcing feature/process

3.2 Dataset Summary

Overall, we have compiled a corpus consisting of 5,584 root words from which a total of 78,629 PoS tagged derived forms were generated. As summarized in Table 1, the nouns constitute 56.73% (44,603 words) of the corpus, while the verbs make up 43.27% (34,026 words). Furthermore, Table 2 shows fine-grained breakdown distributions for the noun and verb subcategories.

Table 1: Dataset distribution by main part of speech

Part of Speech (Qaybta_hadalka)	Count	Percentage
MAGAC (Noun)	44,603	56.73%
FAL (Verb)	34,026	43.27%

4 Pilot Evaluation

In this study, we extend the approach proposed in Mohamed and Mohamed (2023), which introduced a two-stage Somali language lemmatization framework (Figure 3). Briefly, that framework began with the construction of a lexicon by manually compiling and pairing root and derived words based on defined linguistic criteria. In the second stage, written morphological rules were applied to lemmatize words that are not found in the lexicon. This combination of lexicon-based and rule-based methods were employed to pilot test and validate the usability of the lemmatization corpus created in the current work.

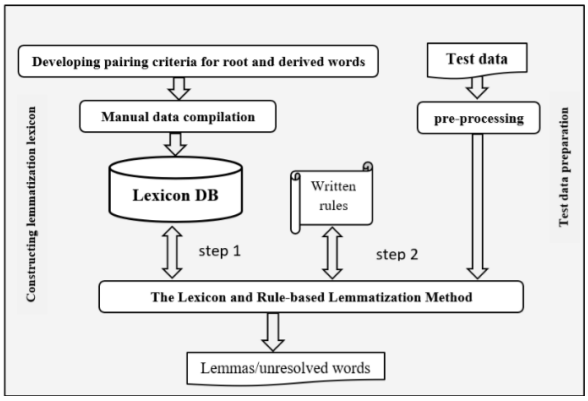

Figure 3: Lexicon and rule-based lemmatization method

Although the present evaluation focuses on rule-based methods, the developed resource is designed to support future neural and LLM-based Somali NLP systems. Large language models typically rely on subword tokenization and often struggle with morphologically rich languages when explicit morphological supervision is absent. The curated

Table 2: Distribution of Grammatical Patterns in the Somali Lexical Dataset

Part of Speech (Qaybta_hadalka)	Count	Percentage
MAGAC+TILMAAN+LAHAANSHO (Noun cloalesced witn possive adjective)	22,955	29.19%
FAL TAAGAN (Progressive verb)	15,115	19.22%
FAL TAGAY (Verb in past form)	15,108	19.21%
MAGAC+TILMAAN+TUSMO (Noun cluster)	6,954	8.84%
MAGAC+QODOB (Defined noun (noun with a definite article))	6,898	8.77%
MAGAC (Noun)	5,587	7.11%
MAGAC+TILMAAN+WEYDIIMO (Noun coalesced with interrogative adjective)	2,209	2.81%
FAL AMAR (Imperetive verb)	1,907	2.43%
FAL MAXADANE (Nonfinite verb)	1,896	2.41%

root–derivative pairs and POS annotations in this dataset can be used to generate supervised training data, construct evaluation benchmarks, guide morphological post-processing, and support hybrid pipelines that combine neural modeling with linguistic constraints. Thus, this work provides foundational linguistic infrastructure that complements and enables future data-driven approaches.

In particular, to evaluate the robustness of the constructed lexicon, 294 Somali text documents of varying lengths were collected, which was a diverse dataset spanning multiple high-frequency public discourse domains of Somali texts sourced from a variety of digital platforms, including social media posts, BBC-Somali, and other reputable online news outlets. The corpus was carefully categorized into 8 thematic domains that reflect the most prominent areas of public discourse in Somali society (Table 3).

Table 3: Distribution of the test dataset by domain

Category	Count	Percentage (%)
Caafimaad (Health)	32	10.88%
Ciyaaro (Sports)	45	15.31%
Diin (Religion)	34	11.56%
Waxbarasho (Education)	20	6.80%
Ganacsi (Business)	68	23.13%
Siyaasad (Political)	62	21.09%
Madadaalo (Entertainment)	14	4.76%
Tiknoolojiyo (Technology)	19	6.46%

Beyond topical diversity, the dataset was further classified according to text size to capture structural and contextual variations across different communication settings. Specifically, the texts were grouped into three categories: small, medium, and long texts (full-length news articles). This stratification is essential for evaluating the lemmatization algorithm, as short texts often exhibit high lexical variability and limited context, whereas long texts present more coherent discourse structures and richer morphological forms.

Quantitative analysis of the dataset revealed that small texts contained an average of 86.56 tokens, medium texts averaged 257.17 tokens, and long texts averaged 563.79 tokens. This clear distinction in token length validates the effectiveness of the size-based categorization and provides a robust foundation for assessing model performance across heterogeneous Somali text types. By incorporating this size-based categorization, the dataset provides a robust foundation for assessing model performance across heterogeneous Somali text types (Figure 4).

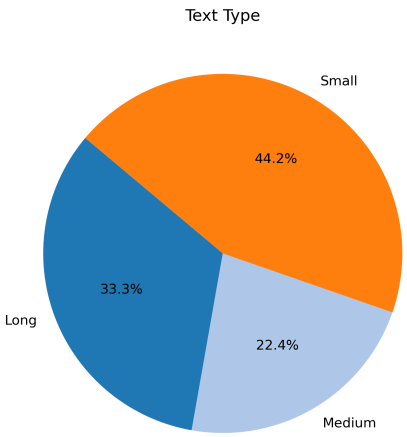

Figure 4: Document token length distributions

Following preprocessing, the corpus was normalized by removing punctuation, numerical tokens, and common stop words, leaving only cleaned textual content for analysis. This step was essential to ensure that morphological processing operated exclusively on linguistically meaningful units. The cleaned dataset was subsequently tested using our rule-based lemmatization system, which integrates lexical resources with Somali-specific morphological rules.

To contextualize the performance of the pro-

posed lexicon-and-rule lemmatization framework, we compare it against two simple baselines. The first baseline is lexicon-only lookup, where tokens are matched directly against the lexical database without applying morphological rules; unmatched tokens remain unresolved. The second baseline is identity mapping, where each token is returned unchanged as its own lemma. These baselines provide lower-bound references that help quantify the contribution of morphological rules beyond dictionary coverage.

The evaluation revealed that the system successfully lemmatized and annotated over 51 percent of the tokens, reducing them to their canonical root forms and simultaneously assigning appropriate part-of-speech (POS) tags. This dual outcome is critical for downstream NLP tasks, as it not only standardizes lexical variation but also provides syntactic and grammatical information that can improve applications such as text classification, machine translation, and information retrieval.

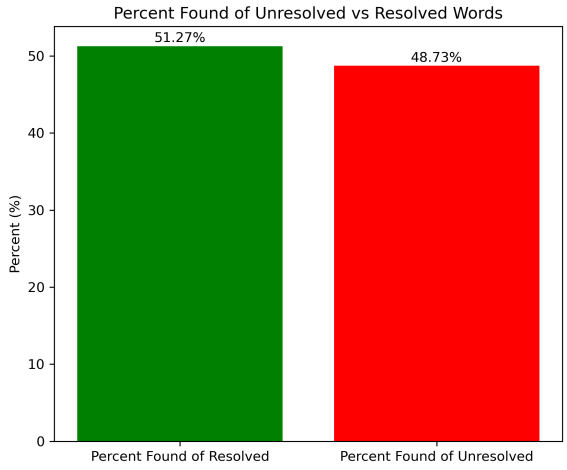

Figure 5: Proportion of tokens successfully lemmatized versus unresolved tokens

Moreover, in addition to domain-level variation, we examined how text size influenced lemmatization outcomes. The results show a clear relationship between text length and system performance. Short texts achieved the highest resolution rate, with approximately 60 percent of tokens assigned a lemma and an associated POS label derived from the lexicon. In contrast, medium-length texts achieved 47% resolution, while long texts recorded the lowest performance.

This trend can be explained by the structural characteristics of different text sizes. Short texts, such as social media posts and user comments, tend to contain fewer tokens and simpler morphological constructions, making them more amenable to rule-based lemmatization. Medium-length texts, while offering richer context, often include greater lexical diversity and more complex derivational structures, which pose challenges to rule-based systems. Long texts, such as full-length news articles and analytical reports, are particularly difficult due to their higher frequency of compounding, derivation, and rare vocabulary, which increases the number of unresolved tokens.

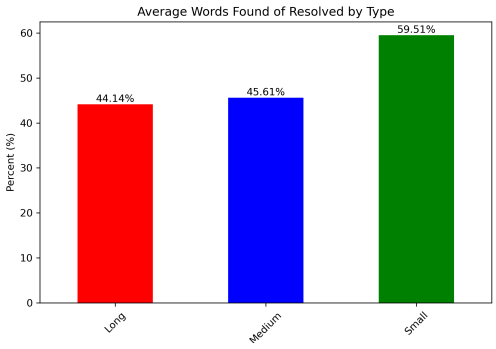

Figure 6: Lemmatization accuracy (percentage of tokens successfully reduced to root forms) across text length categories

To further assess the robustness of the rule-based lemmatizer, we evaluated its performance across different topical domains in the corpus. The results indicate notable variation in resolution rates depending on the domain. Specifically, the average proportion of tokens successfully lemmatized and assigned POS tags was highest in Business (72.38 percent), followed by Education (49.38 percent) and Politics (46.14 percent), while the remaining text categories resolved 45 percent.

This distribution suggests that domains such as business benefited from higher lexical regularity and stronger overlap with the curated lexicon, resulting in a significantly higher resolution rate. In contrast, categories such as technology, health, and sports exhibited lower resolution.

Table 4: % of lemmatised words by text domain

Category	**Words Resolved (%)**
Caafimaad (Health)	43.01%
Ciyaaro (Sports)	43.99%
Diin (Religion)	44.59%
Waxbarasho (Education)	**49.38%**
Ganacsi (Business)	**72.38%**
Siyaasad (Political)	**46.14%**
Madadaalo (Entertainment)	44.67%
Tiknoolojiyo (Technology)	42.38%

5 Discussion

This study demonstrates the application of a linguistically informed and empirically grounded methodology to build a large-scale lemmatization resource for the Somali language with a focus on the MAXAA TIRI written dialect. Our data collection and analysis framework was designed to capture linguistic authenticity, morphological diversity, and domain coverage, all of which are key requirements for evaluating NLP systems in low-resource contexts.

We compiled a total of 5,584 linguistically verified root words sourced from reputable corpus-based platforms and thoroughly validated by Somali linguists to reduce noise and ensure accuracy. These root words were expanded to 78,629 POS-tagged derivatives, creating a comprehensive resource. The resulting lexical database captures the agglutinative and derivational properties of the Somali language which provides a strong foundation for downstream applications, including lemmatization, POS tagging, information retrieval, and classification.

To preserve structural consistency and data integrity, the study introduces a web-based annotation and management platform utilizing relational database architecture with real-time validation to prevent duplication. Controlled crowdsourcing, subjected to expert validation, ensured linguistic precision. Such hybrid annotation strategies increasingly reflect best practices in low-resource NLP, where fully manual annotation is impractical and fully automated methods risk propagating noise. Empirical evaluations across 294 documents spanning multiple textual domains and length categories demonstrated clear performance patterns. Short texts, particularly social media content, achieved the highest lemmatization accuracy (approximately 60%), attributable to simpler syntactic structures and reduced morphological complexity. In contrast, longer news articles demonstrated lower accuracy due to increased lexical variability and complex derivational patterns.

Domain-specific analysis further confirmed that rule-based systems depend heavily on the level of structuredness in the text. Some written business documents reached 72.38% accuracy, benefiting from standardized and repetitive terminology such as 'heshiis' (agreement), 'maalgashi' (investment), 'deyn' (loan), 'dakhliga' (revenue) and 'shirkad' (company), which map cleanly to the constructed lexicon. In contrast, domains such as technology and health showed weaker performance possibly because they contain:

- Borrowed terminology, e.g., 'kombiyuutar' (computer), 'antibaayootik' (antibiotic), etc.
- Code-switching, e.g., 'waxaan update-gareeyay system-ka' (I updated the system), etc.
- Orthographic inconsistency, e.g., 'cafimaad/ caafimaad' (health), etc.

These findings illustrate that dataset diversity, not only in linguistic structure but also in domain and text length, is essential for robust system evaluation. These features challenge deterministic methods by increasing out-of-vocabulary rates and morphological ambiguity. Additionally, the observed performance variation further highlights the limitations of rule-based systems in handling dialectal variation, code-switching, and orthographic inconsistency.

Overall, the study presents a pioneering lemmatization lexicon and a web-based annotation tool that resulted in a linguistically validated Somali NLP resource with its pilot empirical assessment on diverse test datasets showing promising performance. The study's findings also motivate the integration of curated lexical resources with statistical and neural modeling techniques to achieve greater generalization in morphologically complex, low-resource languages.

6 Conclusion

This study developed a scalable and linguistically grounded infrastructure for the Somali NLP resources, moving beyond proof-of-concept to create a reusable lemmatization lexicon. The work delivered a large expert-validated database linking 5,584 roots to over 78,629 POS-tagged derivatives, supported by a web-based annotation and crowdsourcing platform that enables sustainable future data expansion. Empirical testing across 294 documents covering multiple domains and text lengths demonstrated that dictionary and rule-based approach built on the developed dataset can lemmatise more than 51% of tokens overall. More precisely, the performance varied across domains, with the highest accuracy observed in business texts (72%). The system also performed better on shorter and less complex content, such as social media comments.

Future research related to this work will focus on the following directions:

- Expanding the lexical database through controlled crowdsourcing and community-driven annotation to increase domain coverage and scalability.
- Addressing dialectal variation, irregular morphology, orthographic inconsistency, and borrowed terminology through hybrid linguistic–neural approaches.
- Integrating machine learning and deep learning techniques to complement the rule-based system to enable context-sensitive lemmatization.
- Leveraging cross-lingual transfer learning to other morphologically rich languages such as Amharic and Arabic.

Overall, this paper establishes both a novel lexical corpus for Somali and a replicable methodology for other under-resourced languages, demonstrating how linguistic expertise, community participation, and computational methods can converge to build sustainable NLP ecosystems for underrepresented languages.

Data Availability

Datasets are made publicly available for the research community on this GitHub page [2].

References

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, and 26 others. 2022. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

David Ifeoluwa Adelani. 2025. Natural language processing for african languages. *arXiv preprint arXiv:2507.00297*.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure FP Dossou, Akintunde Oladipo, Doreen Nixdorf, and 1 others. 2023. Masakhanews: News topic classification for african languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 144–159.

Jesujoba O Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to african languages via multilingual adaptive fine-tuning. *arXiv preprint arXiv:2204.06487*.

Abdisalam Badel, Ting Zhong, Wenxin Tai, and Fan Zhou. 2023. Somali information retrieval corpus: Bridging the gap between query translation and dedicated language resources. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 7463–7469.

Toms Bergmanis and Sharon Goldwater. 2018. Context sensitive neural lemmatization with lematus. In *16th annual conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1391–1400. Association for Computational Linguistics (ACL).

Cheikh Bamba Dione. 2021. Multilingual dependency parsing for low-resource african languages: Case studies on bambara, wolof, and yoruba. *Proceedings of the International Conference on Parsing Technologies (IWPT)*.

Shai Gordin, Aleksi Sahala, Shahar Spencer, and Stav Klein. 2025. Evacun 2025 shared task: Lemmatization and token prediction in akkadian and sumerian using llms. In *Proceedings of the Second Workshop on Ancient Language Processing*, pages 242–250.

Nizar Habash and Bonnie Dorr. 2003. A categorial variation database for english. Technical report.

Haibin Liu, Tom Christiansen, William A Baumgartner Jr, and Karin Verspoor. 2012. Biolemmatizer: a lemmatization tool for morphological processing of biomedical text. *Journal of biomedical semantics*, 3(1):3.

Enrique Manjavacas, Ákos Kádár, and Mike Kestemont. 2019. Improving lemmatization of non-standard languages with joint learning. *arXiv preprint arXiv:1903.06939*.

Irene Masiringi Mathayo and Alfred Malengo Kondoro. 2024. Unveiling swahili verb conjugations: A comprehensive dataset for low-resource nlp. In *Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval*, pages 149–156.

[2] https://github.com/ShafieAbdi/Somali-Lemmatization-Crowdsourcing-

Aleksandra Miletić and Janine Siewert. 2023. Lemmatization experiments on two low-resourced languages: Low saxon and occitan. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 163–173.

Muhidin A Mohamed, Shuab D Ahmed, Yahye A Isse, Hanad M Mohamed, Fuad M Hassan, and Houssein A Assowe. 2025. Detection of somali-written fake news and toxic messages on the social media using transformer-based language models. *arXiv preprint arXiv:2503.18117*.

Shafie Abdi Mohamed and Muhidin Abdullahi Mohamed. 2023. Lexicon and rule-based word lemmatization approach for the somali language. *arXiv preprint arXiv:2308.01785*.

Siraj Mohammed. 2020. Using machine learning to build pos tagger for under-resourced language: the case of somali. *International Journal of Information Technology*, 12(3):717–729.

Abdillahi Nimaan, Pascal Nocéra, and Jean-François Bonastre. 2006. Towards automatic transcription of somali language. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*.

Antoine Nzeyimana and Andre Niyongabo Rubungo. 2022. Kinyabert: a morphology-aware kinyarwanda language model. *Proceedings of the Association for Computational Linguistics (ACL)*.

Aleksi Sahala, Tero Alstola, Jonathan Valk, and Krister Lindén. 2023. Lemmatizing and pos-tagging akkadian with babylemmatizer and dictionary-based post-correction. In *CLARIN Annual Conference*, pages 111–119. CLARIN ERIC.

Ranka Stanković, Cvetana Krstev, Ivan Obradović, Biljana Lazić, and Aleksandra Trtovac. 2016. Rule-based automatic multi-word term extraction and lemmatization. In *LREC*, pages 507–514.

Jiayi Wang, David Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, and 39 others. 2024. AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced African languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5997–6023, Mexico City, Mexico. Association for Computational Linguistics.

A Appendix

Appendix A: Uploading approved root Words

B Appendix

Appendix B: Annotating root Words

C Appendix

Appendix C: Searching words

D Appendix

Appendix D: Home Page Interface

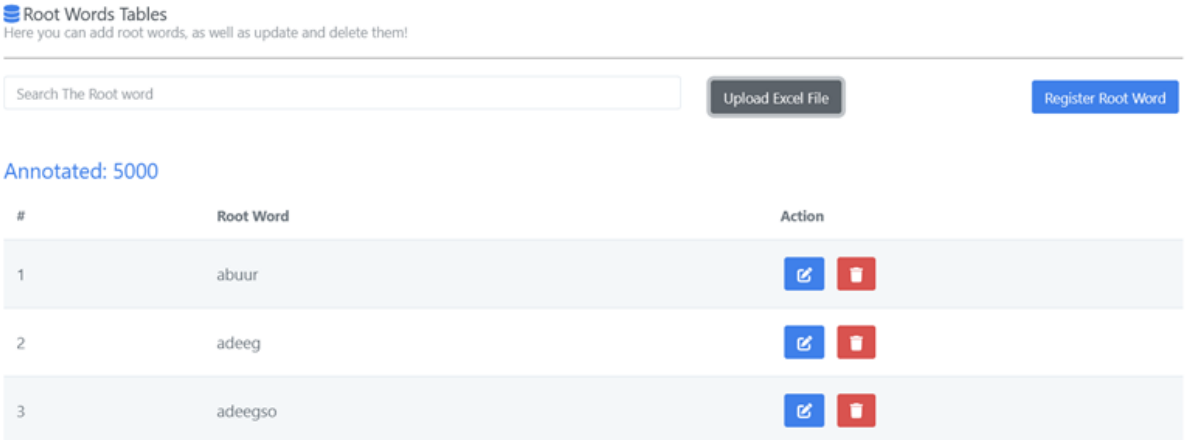

Figure 7: Uploading approved root words to the platform

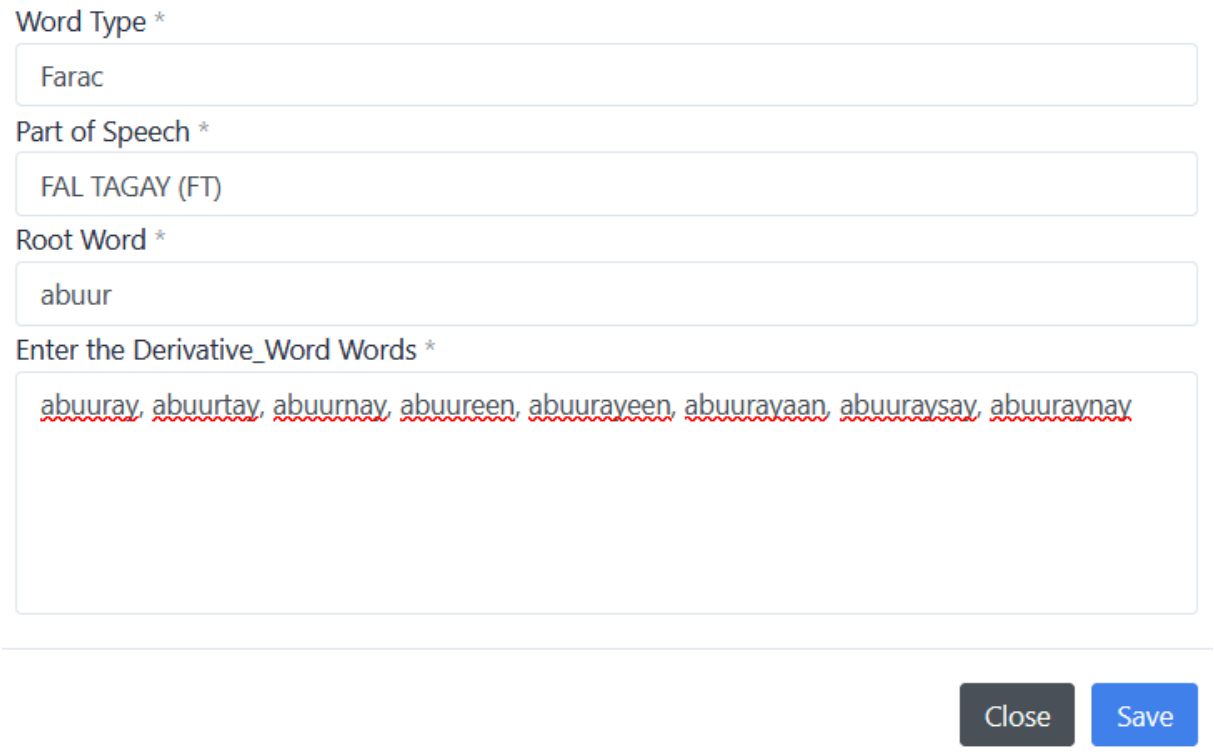

Figure 8: Data Annotation

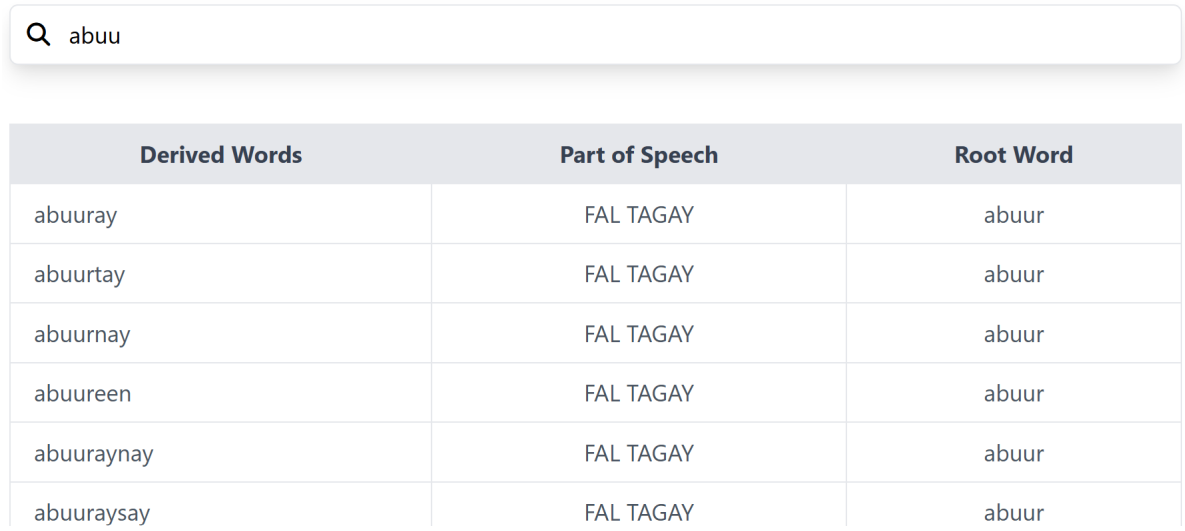

Figure 9: Annotation tool: word search interface

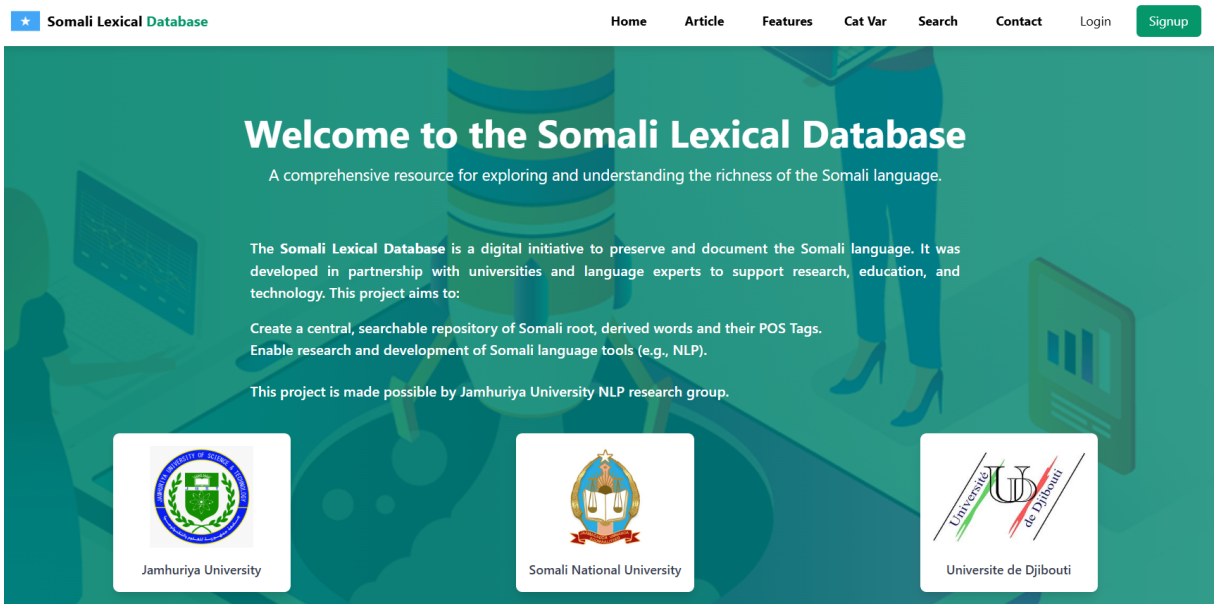

Figure 10: Annotation Tool: Home Page Interface