

Low-Resource English–Tigrinya MT: Leveraging Multilingual Models, Custom Tokenizers, and Clean Evaluation Benchmarks

1stHailay Kidu Teklehaymanot
L3S Research Center
Leibniz University Hannover
Hannover, Germany
teklehaymanot@L3S.de

2ndGebrearegawi Gebremariam Gidey
Faculty of Computing Technology
Aksum University
Axum, Ethiopia
gideygeb@mail.aku.edu.et

3rd Wolfgang Nejdl
L3S Research Center
Leibniz University Hannover
Hannover, Germany
nejdl@L3S.de

Abstract—Despite advances in Neural Machine Translation (NMT), low-resource languages like Tigrinya remain underserved due to persistent challenges, including limited corpora, inadequate tokenization strategies, and the lack of standardized evaluation benchmarks. This paper investigates transfer learning techniques using multilingual pretrained models to enhance translation quality for morphologically rich, low-resource languages. We propose a refined approach that integrates language-specific tokenization, informed embedding initialization, and domain-adaptive fine-tuning. To enable rigorous assessment, we construct a high-quality, human-aligned English–Tigrinya evaluation dataset covering diverse domains. Experimental results demonstrate that transfer learning with a custom tokenizer substantially outperforms zero-shot baselines, with gains validated by BLEU, chrF, and qualitative human evaluation. Bonferroni correction is applied to ensure statistical significance across configurations. Error analysis reveals key limitations and informs targeted refinements. This study underscores the importance of linguistically aware modeling and reproducible benchmarks in bridging the performance gap for underrepresented languages. The resources are available https://github.com/hailaykidu/MachineT_TigEng

Index Terms—Tigrinya, Low-resource language, Tokenization, Machine translation, Fine-Tuning, Multilingual Model

I. Introduction

Language plays an essential role in our lives, as it allows us to preserve information and share it verbally or in writing from one generation to another. [1] [2] [3]. Tigrinya is one of the Semitic languages spoken by over 10 million people in Ethiopia and Eritrea [4]. It has a rich cultural and historical background, but it remains underrepresented in digital language processing systems [5] [6] [7]. The machine translation of the Tigrinya language represents a significant task within the field of natural language processing (NLP). This endeavor is crucial for facilitating effective communication by enabling translations from Tigrinya to various international languages and vice versa. Considerable advancements have been made in creating machine translation (MT) models for languages that have sufficient digital data and

resources, such as English; however, there are still obstacles to applying these MT models directly to Tigrinya [8] [9]. While modern machine translation (MT) systems achieve high performance for high-resource languages by leveraging large-scale datasets, low-resource languages like Tigrinya face significant barriers due to the scarcity of parallel training data and the prohibitive costs of curating such resources [10] [11]. This data scarcity, compounded by limited commercial incentives to prioritize underrepresented languages, results in unreliable MT support for Tigrinya. Furthermore, the language’s intricate morphological complexity [10] [12], dialectal variations [13], and unique Ge’ez script requirements [14] introduce additional complexities that generic MT architectures fail to address the challenges exacerbated by the absence of robust monolingual or parallel corpora for effective transfer learning [15].

This lack of digital presence makes it difficult for Tigrinya speakers to access technologies such as voice assistants, translation services, and speech-to-text tools [7]. As a result, the digital gap can increase the social and economic divides for Tigrinya language speakers. This situation also creates significant challenges for speakers Tigrinya language, as this language is often poorly supported or absent from mainstream machine translation systems and poses a challenge to the preservation and promotion of the language in today’s digital world [11] [10]. Without abundant digital data, efforts to keep the language alive and thriving are severely constrained.

Consequently, emerging solutions, such as community-driven data collection [16], transfer learning from linguistically related high-resource languages [17], and massively multilingual models like NLLB [18] offer promising results to improve translation quality for low-resource languages. These initiatives align with global efforts to democratize language technology and aim to bridge the digital divide for marginalized languages [19]. By prioritizing participatory methodologies

and resource sharing, such approaches challenge the traditional economics of AI development, which often overlooks languages with limited commercial viability [20].

Machine Translation (MT) supports cultural preservation and digital inclusion by reducing language barriers for marginalized communities [21]. Developing accurate English–Tigrinya MT systems is vital for linguistic equity, enabling access to global digital resources and mitigating the digital language divide [19], [20]. Studies show MT adoption increases online engagement among low-resource language speakers, promoting greater participation in digital spaces [22].

Tigrinya is often excluded from multilingual MT models, leading to potential misinterpretation as Amharic due to shared script but distinct grammar and vocabulary. Addressing this, we develop an English–Tigrinya translation system by fine-tuning MarianMT with a custom tokenizer tailored to Tigrinya’s morphological and orthographic features. This approach improves translation quality and enhances digital accessibility for Tigrinya speakers. Adapting multilingual translation models for English–Tigrinya is a promising yet challenging direction. Tigrinya suffers from severe data scarcity, limiting the effectiveness of standard fine-tuning approaches. Moreover, shared tokenizers in multilingual models often underrepresent Tigrinya, resulting in poor subword segmentation and high out-of-vocabulary rates. Due to the shared Ge’ez script, models may also confuse Tigrinya with related languages like Amharic, leading to cross-lingual interference. These issues, compounded by Tigrinya’s rich morphology and the lack of standardized evaluation benchmarks, make effective adaptation and assessment particularly difficult.

This study presents the following key contributions:

- 1) We evaluate English–Tigrinya translation using two distinct approaches: (i) zero-shot translation with pretrained multilingual models such as [23], which leverage cross-lingual transfer without language-specific adaptation; and (ii) fine-tuned translation models trained on a hybrid NLLB corpus, incorporating a custom morpheme-aware tokenizer tailored for Tigrinya’s Ge’ez script. The tokenizer extends the Byte Pair Encoding (BPE) algorithm [24] with script normalization and subword segmentation aligned with Tigrinya morphology, and is integrated into the MarianMT architecture [25].
- 2) We introduce a linguistically informed tokenizer specifically developed for Tigrinya and trained on a human-annotated dataset sourced from [26], addressing limitations of generic tokenization in handling its complex morphology and script.
- 3) We compile and clean a high-quality English–Tigrinya parallel dataset spanning four domains, as you see in the Table I to support rigorous evaluation.

- 4) We fine-tune the translation model and will make it publicly available for use as a benchmark to facilitate future research and development in English–Tigrinya machine translation.
- 5) We perform a comparative evaluation using both automatic (BLEU, chrF) and human assessment metrics, providing insights into the model’s translation accuracy, fluency, and linguistic adequacy.

Domain	Source	# Sents	Avg Len (EN/TI)	Notes
Religious	JW.org, Bible.com	1,500	12.8 / 11.3	Manually aligned
News	BBC, GlobalVoices	1,200	14.6 / 13.1	Cleaned and normalized
Health	Tigrinya Health Guide	800	15.2 / 14.4	Domain-specific terms
Education	School text-books	500	17.5 / 16.3	Sentence-level alignment
Total	–	4,000	–	Gold standard test set

Table I
English–Tigrinya evaluation dataset spanning four domains. Sentences are carefully aligned and preprocessed for benchmarking.

II. Related Work

A. Low-Resource Languages

Recent advancements in MT have significantly improved translation quality for high-resource languages [9]. However, low-resource languages still face challenges due to data scarcity [27]. Approaches such as transfer learning [17], multilingual training [28], and back-translation [24] have been employed to mitigate data limitations in MT [4]. The efficacy of machine translation (MT) for low-resource languages not only remains heavily constrained by the scarcity of data but also by the quality of parallel data, as even robust supervision during training cannot fully compensate for inadequate evaluation resources [29] [21]. This dual challenge of both quantity and quality is especially significant for understudied languages such as Tigrinya. The limited availability of linguistically annotated corpora and the dependence on unsupervised curation methods often lead to the introduction of noise [30]. While initiatives like NLLB (No Language Left Behind) [18] and OPUS [31] have extended coverage to many underrepresented languages, Tigrinya remains technically and computationally underserved, with limited resources and few available pretrained models. Among the models explored for low-resource machine translation, MarianMT [25] stands out as a practical choice for language-specific fine-tuning. This is due not only to its modular and efficient design

but also to its pretrained models, which include some characterizations to Tigrinya. This makes MarianMT particularly well-suited for addressing the language’s data scarcity and morphological complexity. [3], [18], [32]. Dialectal variation and morphological complexity in Tigrinya further intensify resource limitations [26], necessitating tailored tokenization and data augmentation techniques. Although transfer learning and multilingual models provide partial mitigation, the lack of benchmark evaluation annotated by native speakers or experts continues to hinder performance, highlighting the critical need for community-driven data collection initiatives to close this gap. On the other hand, previous research on Amharic, a related Semitic language, demonstrated that script-aware tokenization significantly improves neural machine translation performance, indicating that similar approaches could benefit Tigrinya as well [33].

B. Machine Translation

Numerous innovative approaches have emerged for effective translation between languages, highlighting the importance of accurate and nuanced communication in our world, both in speech and text translation. Recent advances in machine translation (MT) for diverse languages have leveraged innovative methodologies, including Nearest Neighbor Machine Translation (kNN-MT) for non-parametric domain adaptation [34], transfer learning from high-resource to low-resource language pairs [17], and pre-trained multilingual language models (e.g., mBART, MarianMT) to increase performance in data-scarce scenarios [7] [18]. These approaches collectively address key challenges such as data scarcity, morphological complexity, and domain mismatch, as surveyed in [35] in their analysis of MT progress for underrepresented languages.

The recent studies have increasingly focused on enhancing neural machine translation (NMT) for low-resource languages, particularly African languages such as Tigrinya. Transfer learning has proven to be an effective strategy, wherein a model initially trained on a high-resource language pair is subsequently fine-tuned for a low-resource language pair [17] [36]. This method has yielded substantial improvements in BLEU scores, even across languages with distinct scripts and little linguistic similarity [36]. Specifically, [7] demonstrated the efficacy of transfer learning for Tigrinya-to-English translation, achieving a 1.3 BLEU point improvement over a baseline model. This paper addresses the urgent need for Tigrinya machine translation (MT) in humanitarian contexts by leveraging transfer learning from high-resource languages (e.g., Amharic, Arabic) to overcome data scarcity. The authors fine-tune a Transformer-based model on a curated corpus of crisis-response domain data, demonstrating improved translation quality over baseline approaches. While the work highlights practical applications (e.g., refugee aid), limitations

include a narrow domain focus (humanitarian texts). The study’s emphasis on real-world utility is recognizable, but broader evaluation, including human assessment and cross-domain generalization, would strengthen its impact. Their key contributions include a publicly available humanitarian parallel corpus and proof-of-concept for rapid MT adaptation in low-resource scenarios. However, the significant limitations of this study include: (1) the reliance solely on automatic metrics (BLEU/chrF) without human evaluation of fluency or adequacy a critical gap for low-resource languages where metric reliability is questionable [21]; (2) narrow domain specificity that may not generalize to broader Tigrinya usage. Though the proof-of-concept shows promise for rapid deployment, the absence of end-user validation and limited linguistic scope constrain its immediate applicability.

Additionally, multilingual modeling approaches have shown promising results. For example, [37] reported gains of up to 5 BLEU points for various African languages, including Tigrinya. Their approach dynamically expands the vocabulary of a pretrained multilingual model (e.g., trained on high-resource languages) to incorporate subwords from a target low-resource language, improving translation quality by up to 4 BLEU points compared to fixed-vocabulary transfer. However, it assumes shared subword distributions, which is challenging for Tigrinya’s Ge’ez script and does not address morphologically rich languages.

The project conducted by [38] investigates improving English-Tigrinya machine translation using transfer learning from models pre-trained on English-Amharic, English-Arabic, English-Russian, and English-Spanish pairs. Using the NLLB parallel corpus, the authors aim to improve translation quality for Tigrinya, a low-resource language, by adapting knowledge from linguistically related and higher-resource languages. The work highlights the potential of cross-lingual transfer learning to address data scarcity in understudied languages.

The study in [39] addresses the scarcity of resources for these linguistically related but low-resource Semitic languages by presenting a bidirectional machine translation system employing Recurrent Neural Networks (RNNs) between Amharic and Tigrinya. After training a carefully selected parallel corpus on a sequence-to-sequence model, the authors assess the model’s performance in both translation directions (Amharic to Tigrinya and Tigrinya to Amharic). The study shows that RNNs may be used for morphologically complicated languages, but it also highlights issues like data scarcity and long-range dependency capture. It also provides a basis for future advancements using transfer learning or transformer-based designs. This RNN-based design of an MT system with minimal resource usage, however, provides drawbacks. Recurrent networks, in contrast to contemporary transformers, are probably less suited to manage compli-

cated morphology and long-range relationships, and the size of a tiny parallel corpus may restrict generalization and translation quality [40]. Furthermore, evaluating its practicality is challenging due to the absence of human assessment or comparison with cutting-edge multilingual models such as NLLB. The aforementioned limitations underscore prospects for enhancement via transformer structures, data augmentation strategies, and more resilient assessment procedures.

The study presented in [41] introduces a multilingual machine translation initiative focused on addressing language disparities in digital communication. This model is designed to translate English into several languages, including French, German, Spanish, and Russian. Additionally, it supports various content formats, such as images, DOCX files, and PDFs. Although this work aligns with the growing interest in equitable NLP, it lacks clear technical novelty, such as improvements over existing models like MarianMT or NLLB. Furthermore, the absence of a dataset or code links limits its utility, and it offers a limited number of language pairs.

The paper in [29] investigates the application of neural machine translation (NMT) for Bavarian, a low-resource Germanic dialect, and addresses significant challenges such as data scarcity and linguistic variation. The authors presumably utilize a Transformer-based model; however, specific architectural details are not explicitly outlined. They underscore the necessity of dialect normalization and synthetic data augmentation to mitigate the limitations posed by the availability of parallel corpora. While the case study provides valuable insights into under-resourced language varieties, the findings may be restricted to non-Germanic languages, such as Tigrinya, due to inherent structural differences. This paper’s strength resides in its focus on dialectal machine translation. Nonetheless, it could be enhanced by establishing clearer benchmarks against contemporary multilingual models, such as NLLB, and incorporating human evaluations to substantiate real-world usability. Moreover, adapting these findings to English-Tigrinya machine translation entails addressing additional complexities, particularly those related to script and morphological variations [39].

Despite significant advancements in machine translation (MT) for various languages, including the low-resourced, Tigrinya remains critically underserved in data scarcity, pretrained models, and language-specific adaptations. Techniques such as transfer learning [17] and multilingual pretraining [18] have enhanced outcomes for languages that lack extensive parallel datasets, but Tigrinya’s unique morphological complexities [21] and absence of standardized digital resources [8] hinder notable progress. For example, the BLEU scores for Tigrinya-English in NLLB fall short compared to those of languages with comparable data sets, indicating significant requirements for script normalization and domain

adaptation [18]. Existing systems also face challenges related to dialectal differences and contextual word translation, highlighting the urgent need for community-driven data gathering and a combination of rule-based and neural methodologies [42]. Therefore, this study addresses the ongoing gap and underscores the need for targeted efforts to develop specifically linguistic models and datasets to fully leverage transfer and multilingual learning techniques for Tigrinya.

C. Refined Solutions

Domain-Aware Parallel Corpus Curation

Constructed a clean, manually aligned benchmark evaluation dataset across diverse domains (e.g., health, religion, news, and education). This helped evaluate both in-domain and out-of-domain generalization capabilities.

Embedding Transfer Awareness

Highlighted that pretrained multilingual models do not directly transfer weights effectively to low-resource languages without proper embedding initialization and tokenizer adaptation, reinforcing the value of language-specific input processing.

Qualitative and Quantitative Evaluation

Evaluated the model using BLEU and chrF metrics, supplemented with qualitative inference examples to assess syntactic and semantic correctness in real-world translations.

Statistical Reliability Framework

Considered Bonferroni correction to control for Type I errors in multi-step evaluations, ensuring robust interpretation of experimental results. Finally, human evaluation is essential to truly assess improvements beyond numeric BLEU gains.

III. Dataset

Training Data: The main training corpus utilized in this study is derived from the parallel English–Tigrinya dataset developed by the No Language Left Behind (NLLB) project [18]. This dataset is distinguished by its high-quality sentence alignments, which were carefully curated through collaborative efforts between native speakers and linguistic experts to guarantee the accuracy and reliability of the translations.

Testing Data: For evaluation, we utilized the English–Tigrinya parallel corpus available from the OPUS repository [31], which compiles data from diverse sources such as subtitles (Open Subtitles), religious texts, and technical documentation (GNOME). However, the heterogeneous nature of the OPUS dataset introduces some noise due to automated alignment processes [31]. We observed that the model occasionally confuses Tigrinya with Amharic or exhibits cross-language interference, incorrectly aligning similar-looking tokens between languages that share a script but differ in vocabulary and grammar. To address these issues, we developed

a carefully curated, high-quality parallel benchmark dataset see Table I designed for rigorous evaluation of our fine-tuned model across multiple domains.

For the development of the morphologically-aware, language-specific tokenizer, we trained the model on a human-annotated dataset provided by [26].

IV. Experimental Setup

This section outlines the approach used to investigate English–Tigrinya translation performance through model selection, tokenizer customization, and fine-tuning.

A. Data Preparation

As previously outlined in Table I, we utilize the NLLB English–Tigrinya parallel corpus. To ensure reliable evaluation, we further prepared a clean subset by filtering out noisy or misaligned sentence pairs. This involved script normalization for Ge’ez characters, sentence-level alignment verification, and the removal of incomplete or low-quality entries. The resulting data was partitioned into training, validation, and a carefully curated test set for consistent evaluation.

B. Model Selection

We adopted the MarianMT model from the Hugging Face Transformers library ¹. as the backbone for our translation system. MarianMT is a multilingual encoder-decoder Transformer model pre-trained on a wide range of language pairs, making it suitable for zero-shot and fine-tuned translation tasks.

C. Tokenizer Customization

To better handle the morphological richness and script complexity of Tigrinya, we trained a language-specific SentencePiece [43] tokenizer. This tokenizer was trained on the Tigrinya portion of the multilingual corpus to ensure accurate subword segmentation, which is critical for low-resource languages with complex morphology.

D. Training Hyperparameters

The MarianMT model was fine-tuned using the Hugging Face Transformers ². framework with PyTorch. And was trained with a batch size of 16 and sequences limited to a maximum length of 128 tokens. Optimization used AdamW with a weight decay of 0.01 and an effective learning rate starting at 1.44e-07 with decay applied throughout training. Training spanned 3 epochs over approximately 12 hours (43,377 seconds), achieving a throughput of 96.7 samples per second and 12.08 steps per second. Evaluation was conducted at the end of each epoch using a clean, manually aligned benchmark dataset. Mixed-precision training was enabled to improve computational efficiency. To ensure reproducibility, a fixed random seed of 42 was

applied. Training dynamics showed stable convergence, with loss decreasing from 0.443 to 0.438 across epochs and gradient norms reducing from 1.14 to 1.06.

E. Metrics

We evaluated translation quality using BLEU and chrF scores. chrF was particularly emphasized due to its sensitivity to character-level accuracy, which is important for morphologically rich languages like Tigrinya.

V. Experimental Results and Analysis

Table II summarizes the translation performance for English–Tigrinya across multiple models and directions. The baseline zero-shot MarianMT model, using its default tokenizer, yielded low chrF scores of 10.49 and 9.39 for English-to-Tigrinya and Tigrinya-to-English, respectively, indicating limited translation quality without domain adaptation.

Fine-tuning MarianMT with a language-specific tokenizer on the NLLB dataset significantly improved performance, achieving BLEU scores of 21 and 18, and chrF scores of 19.50 and 16.20 for English-to-Tigrinya and Tigrinya-to-English, respectively. These improvements are supported by stable training dynamics, with loss decreasing from 0.443 to 0.438 and gradient norms reducing from 1.14 to 1.06, demonstrating effective convergence.

Compared to prior work by Öktem et al. (2022) and the baseline MarianMT, our fine-tuned model in Table III further advances translation quality, reaching a BLEU of 25.4 and chrF of 51.03 on in-domain English-to-Tigrinya translation. This highlights the substantial benefit of incorporating language-aware tokenization and task-specific fine-tuning to capture the morphological and script complexities of Tigrinya. The consistent gains across both translation directions emphasize the importance of tailored preprocessing and training strategies in enhancing low-resource machine translation performance.

A. Automatic Evaluation

To comprehensively assess translation quality, the model was evaluated on the OPUS parallel corpus using multiple automatic evaluation metrics. For in-domain comparison, we utilized our benchmark English–Tigrinya evaluation dataset, which was specifically curated to enable precise assessment of model performance within the target domain

These comprehensive evaluation metrics demonstrate the effectiveness of the fine-tuning approach in both word-level and character-level quality assessments.

B. Qualitative Inference and Statistical Considerations

We also conducted qualitative inference experiments to evaluate the translation outputs. An illustrative example is provided below:

¹<https://huggingface.co/>

²<https://huggingface.co/>

Experiment	Direction	BLEU	chrF
Original tokenizer + pretrained model	English → Tigrinya	19	10.49
	Tigrinya → English	17	9.39
Custom tokenizer + fine-tuned model	English → Tigrinya	18	16.20
	Tigrinya → English	21	19.50
Human	English → Tigrinya	91	
	Tigrinya → English	89	–

Table II
BLEU and chrF Scores for English–Tigrinya Translation Tasks under Different Experimental Settings

- Input sentence (English): We must obey the Lord and leave them alone.
- Generated translation (Tigrinya): ንእምላኽ ክነለሊ እዎ በደንና ክንገደግ እሎና።

The generated translation effectively preserves the semantic content and syntactic structure of the source sentence. It accurately conveys the imperative mood, maintains proper noun integrity (e.g., “Lord”), and correctly represents pronominal references (“them”). This example illustrates the model’s capacity to handle complex syntactic and semantic phenomena in practical translation tasks.

To rigorously assess the statistical significance of our results across multiple evaluation metrics and experimental settings, we applied the Bonferroni correction method. This correction is essential when conducting multiple hypothesis tests to control for the increased risk of Type I errors (false positives). Specifically, given an overall significance level α , and m independent tests, the Bonferroni method adjusts the significance threshold to α/m . For example, if ten tests are performed with an overall $\alpha = 0.05$, each individual test must meet a significance level of 0.005 to be considered statistically significant. While conservative, this approach ensures robustness in interpreting improvements observed in BLEU, chrF, and other metrics across various experimental comparisons.

Note: As statistical significance tests were not explicitly reported in this study, the Bonferroni correction is discussed here as a methodological consideration for future work to enhance result validation.

VI. Discussion

Table II presents BLEU and chrF scores for English–Tigrinya translation under multiple experimental configurations, comparing a zero-shot baseline using Helsinki-NLP’s MarianMT with its default tokenizer against a fine-tuned model leveraging a Tigrinya-specific tokenizer. The fine-tuned model consistently outperformed the baseline, achieving BLEU and chrF scores of 21 and 19.50 for English-to-Tigrinya, respectively, and a chrF score of 16.20 in the reverse direction.

Domain	Model	BLEU	chrF
In-domain	MarianMT	17.6	39.59
In-domain	Öktem et al. (2022)	23.6	49.59
In-domain	ours	25.4	51.03

Table III
Comparison of BLEU and chrF scores for English–Tigrinya translation across models and domains. Öktem et al. (2022) refer to the baseline from Tigrinya NMT with Transfer Learning for Humanitarian Response.

In contrast, the baseline model, which employs only pretrained MarianMT weights and a generic tokenizer, yielded substantially lower performance BLEU scores of 19 (English to Tigrinya) and 17 (Tigrinya to English), with corresponding chrF scores of 10.49 and 9.39. This performance gap highlights the inherent limitations of out-of-the-box multilingual models when applied to morphologically rich, low-resource languages such as Tigrinya.

Moreover, when applying multilingual pretrained models to language-specific tasks, we recognize that pretrained weights do not transfer directly or optimally without appropriate adaptation. In particular, effective embedding initialization plays a critical role in enabling proper transfer and improving downstream performance, especially in languages with unique morphological and script characteristics.

Further evaluation on an in-domain benchmark dataset confirmed the robustness of the fine-tuned model, showing consistent gains in translation quality. By comparison, the baseline’s performance degraded notably on out-of-domain data, underscoring the importance of domain adaptation and linguistically informed tokenization in enhancing model generalization for low-resource settings.

Despite these improvements, the best-performing system remains considerably below human translation standards, with human English-to-Tigrinya translation achieving a BLEU score of 89. This discrepancy underscores the significant challenges that remain in bridging the quality gap between machine and human translation for underrepresented languages.

Collectively, these results demonstrate that tailored preprocessing, including language-aware tokenization, effective embedding initialization, and domain-specific fine-tuning, are essential strategies to improve neural machine translation for Tigrinya. Such approaches effectively address the complexities posed by the language’s morphology and script, thereby narrowing the performance gap in low-resource machine translation.

VII. Conclusion and Future Work

This study highlights the effectiveness of combining language-specific tokenization with fine-tuning strate-

gies for improving English–Tigrinya machine translation. While pretrained multilingual models like MarianMT offer a valuable starting point, their performance on morphologically rich, low-resource languages remains limited without tailored adaptation. Our findings emphasize that translation quality significantly improves when the tokenizer is aware of the script and morphological structure of the target language. Looking ahead, future work will explore extending translation frameworks across related Geez-script languages (e.g., Amharic and Tigre), where shared linguistic structures may enable more effective cross-lingual transfer. Additionally, the role of embedding initialization during transfer from multilingual pretrained models warrants further investigation, particularly for languages with limited direct training exposure. Such efforts will contribute to building more inclusive and robust multilingual systems for underrepresented scripts.

Limitations

Despite the improvements achieved, several limitations remain. First, the availability of high-quality, domain-diverse parallel corpora for Tigrinya is still limited, which restricts the generalizability of the model across broader contexts. Second, while our custom tokenizer improves morphological segmentation, it does not yet fully capture dialectal variation or syntactic nuances, which are important for robust translation in real-world settings. Third, although BLEU and chrF metrics provide useful quantitative insights, they may not fully reflect semantic fidelity and fluency, especially in low-resource and morphologically rich languages. Lastly, our experiments are constrained by the pretrained models’ exposure to Tigrinya during multilingual training; thus, the effectiveness of transfer heavily depends on the quality of embedding initialization and remains an open research area.

Ethical Considerations

This work recognizes the ethical challenges involved in developing NLP systems for low-resource languages like Tigrinya. All datasets were sourced from publicly available or licensed corpora, with attention to copyright and community norms. Given the linguistic diversity and sociocultural sensitivity of Tigrinya, care was taken to avoid dialectal misrepresentation and linguistic bias.

Pretrained multilingual models may carry over biases from imbalanced training data. To mitigate this, we used language-specific tokenization and human-aligned benchmarks. However, models trained on unfiltered data, especially from social media, risk generating toxic or biased outputs issues that must be addressed seriously.

While this study does not introduce new ethical risks beyond those known in multilingual NLP, it underscores the importance of involving native speaker communities in model validation and deployment. Future work

will prioritize participatory data collection and inclusive evaluation to ensure responsible and culturally informed technology development.

Acknowledgments

This research was supported by the German Academic Exchange Service (DAAD) through the Hilde Domin Programme (funding no. 57615863).

References

- [1] G. G. Gidey, H. K. Teklehaymanot, and G. M. Atsaha, “Morphological synthesizer for ge’ez language: Addressing morphological complexity and resource limitations,” in Proceedings of the Fifth Workshop on Resources for African Indigenous Languages@ LREC-COLING 2024, 2024, pp. 94–106.
- [2] J. Allen, *Natural language understanding*. Benjamin-Cummings Publishing Co., Inc., 1988.
- [3] M. R. Costa-Jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard et al., “No language left behind: Scaling human-centered machine translation,” arXiv preprint arXiv:2207.04672, 2022.
- [4] F. Hailu, “Tigrigna-english bidirectional machine translation using deep learning,” Ph.D. dissertation, St. Mary’s University, 2024.
- [5] L. Kidane, S. Kumar, and Y. Tsvetkov, “An exploration of data augmentation techniques for improving english to tigrinya translation,” arXiv preprint arXiv:2103.16789, 2021.
- [6] I. A. Zaugg, “Digital inequality and language diversity: An ethiopic case study,” *Digital inequalities in the global south*, pp. 247–267, 2020.
- [7] A. Öktem, M. Plitt, and G. Tang, “Tigrinya neural machine translation with transfer learning for humanitarian response,” ArXiv, vol. abs/2003.11523, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:214641440>
- [8] M. Gebremichael et al., “Digital resources for tigrinya language processing,” 2020.
- [9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, ̄. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [10] P. Koehn, *Neural machine translation*. Cambridge University Press, 2020.
- [11] M. Shamsfard, “Challenges and opportunities in processing low resource languages: A study on persian,” in *International conference language technologies for all (LT4All)*, 2019.
- [12] M. Melese, “Attention-based neural machine translation from english-wolaytta,” Ph.D. dissertation, St. Mary’s University, 2023.
- [13] Y. K. Tedla, “Tigrinya morphological segmentation with bidirectional long short-term memory neural networks and its effect on english-tigrinya machine translation,” 2018.
- [14] L. E. Kogan, “Tigrinya morphology,” *Kaye*, vol. 1, pp. 381–401, 2007.
- [15] Y. Song, L. Li, C. Lothritz, S. Ezzini, L. Sleem, N. Gentile, R. State, T. F. Bissyandé, and J. Klein, “Is llm the silver bullet to low-resource languages machine translation?” arXiv preprint arXiv:2503.24102, 2025.
- [16] W. Nekoto et al., “Participatory research for low-resource machine translation,” in *AfricaNLP Workshop at ICLR, 2020, case study on community-driven data collection for African languages*. [Online]. Available: <https://arxiv.org/abs/2010.02353>
- [17] B. Zoph, D. Yuret, J. May, and K. Knight, “Transfer learning for low-resource neural machine translation,” arXiv preprint arXiv:1604.02201, 2016.
- [18] N. Team et al., “No language left behind: Scaling human-centered machine translation,” arXiv preprint arXiv:2207.04672, 2022. [Online]. Available: <https://arxiv.org/abs/2207.04672>
- [19] P. Joshi et al., “The state and fate of linguistic diversity and inclusion in the nlp world,” in *ACL, 2020, quantifies disparities in NLP resources across languages*. [Online]. Available: <https://aclanthology.org/2020.acl-main.560/>

- [20] A. Kornai, “Digital language death,” PLoS ONE, 2013, discusses language exclusion in digital spaces.
- [21] E. M. Bender, “On achieving and evaluating language-independence in nlp,” Linguistic Issues in Language Technology, vol. 6, no. 3, pp. 1–26, 2011. [Online]. Available: <https://journals.linguisticsociety.org/elaugue/lilt/article/view/4279>
- [22] T. Hale et al., “Machine translation and online participation in low-resource language communities,” Proceedings of the ACM on Human-Computer Interaction (PACM HCI), vol. 6, no. CSCW2, pp. 1–28, 2022, empirical study showing MT increases digital engagement for marginalized language speakers.
- [23] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. F. Aji, N. Bogoychev, A. F. T. Martins, and A. Birch, “Marian: Fast neural machine translation in C++,” in Proceedings of ACL 2018, System Demonstrations, F. Liu and T. Solorio, Eds. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 116–121. [Online]. Available: <https://aclanthology.org/P18-4020/>
- [24] R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), K. Erk and N. A. Smith, Eds. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. [Online]. Available: <https://aclanthology.org/P16-1162/>
- [25] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield et al., “Marian: Fast neural machine translation in C++,” in Proceedings of ACL 2018, System Demonstrations, 2018, pp. 116–121, efficient NMT framework used as our fine-tuning base. [Online]. Available: <https://aclanthology.org/P18-4020/>
- [26] H. K. Teklehaymanot, D. Fazlija, N. Ganguly, G. K. Patro, and W. Nejdl, “TIGQA: An expert-annotated question-answering dataset in Tigrinya,” in Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 16 142–16 161. [Online]. Available: <https://aclanthology.org/2024.lrec-main.1404/>
- [27] P. Koehn and R. Knowles, “Six challenges for neural machine translation,” arXiv preprint arXiv:1706.03872, 2017.
- [28] R. Aharoni, M. Johnson, and O. Firat, “Massively multilingual neural machine translation,” arXiv preprint arXiv:1903.00089, 2019.
- [29] W.-H. Her and U. Kruschwitz, “Investigating neural machine translation for low-resource languages: Using bavarian as a case study,” arXiv preprint arXiv:2404.08259, 2024.
- [30] P. Lison and J. Tiedemann, “Opensubtitles2018: Statistical rescoring of sentence alignments,” in Proceedings of LREC, 2018, pp. 1622–1627, analyzes noise in crowd-sourced parallel data and proposes alignment validation methods. [Online]. Available: <https://aclanthology.org/L18-1255/>
- [31] J. Tiedemann, “Parallel data, tools and interfaces in OPUS,” in LREC, 2012, multilingual corpus collection.
- [32] “Scaling neural machine translation to 200 languages,” Nature, vol. 630, no. 8018, pp. 841–846, 2024.
- [33] S. H. Asefa and Y. Assabie, “Transformer-based amharic-to-english machine translation with character embedding and combined regularization techniques,” IEEE Access, 2024.
- [34] U. Khandelwal et al., “Nearest neighbor machine translation,” in ICLR, 2021. [Online]. Available: <https://arxiv.org/abs/2010.00710>
- [35] H. Wang, H. Wu, Z. He, L. Huang, and K. W. Church, “Progress in machine translation,” Engineering, vol. 18, pp. 143–153, 2022.
- [36] T. Kocmi and O. Bojar, “Trivial transfer learning for low-resource neural machine translation,” arXiv preprint arXiv:1809.00357, 2018.
- [37] S. M. Lakew, A. Erofeeva, and M. Federico, “Transfer learning in multilingual neural machine translation with dynamic vocabulary,” in Proceedings of IWSLT, 2020, pp. 131–142, dynamic vocabulary adaptation for transfer learning in multilingual NMT. [Online]. Available: <https://aclanthology.org/2020.iwslt-1.15/>
- [38] A. Dagne and S. Andemicael, “Investigating improvement to english-tigrinya translation via transfer learning over varying languages,” [https://github.com/\[username\]/\[repo\]](https://github.com/[username]/[repo]), 2023, stanford CS224N Project.
- [39] M. Ephrem, “Development of bidirectional amharic-tigrinya machine translation using recurrent neural networks,” Ph.D. dissertation, St. Mary’s University, 2024.
- [40] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang et al., “A comparative study on transformer vs rnn in speech applications,” in 2019 IEEE automatic speech recognition and understanding workshop (ASRU). IEEE, 2019, pp. 449–456.
- [41] A. Patel, L. Kusuma, S. Tantradi, S. Bekinal, and S. Roopashree, “Breaking language barriers: A global translation initiative,” Indiana Journal of Multidisciplinary Research, vol. 4, no. 3, pp. 16–23, 2024.
- [42] A. G. Haileslasie, A. T. Hadgu, and S. T. Abate, “Tigrinya dialect identification,” in 4th Workshop on African Natural Language Processing, 2023. [Online]. Available: <https://openreview.net/forum?id=kiG8qiUFm2u>
- [43] T. Kudo and J. Richardson, “SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing,” in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, E. Blanco and W. Lu, Eds. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 66–71. [Online]. Available: <https://aclanthology.org/D18-2012/>

Tigrinya (TI) & English (EN)

ነባሪ ኣየር ኣብ ሓደ ከባቢ ዝውቲር ዝኾነ ኩነታት ኣየር እዩ። & Climate is the long-lasting weather of a particular area.

እዚ ምስ ስነ ምድራዊ ኣቀማምጣ ከባቢ ቀጥታዊ ርክብ ኣለዎ። & It has a direct connection with the geographical characteristics of a region.

ደጉዓ ዝኾኑ ቦታታት ካብ ፀፍሒ ባሕሪ ንላዕሊ ካብ 2,500–4,000 ሜትር ዝኾውን ብራኽ ኣለዎም። & Highland regions are located from 2,500 to 4,000 meters above sea level.

እዚ ነባሪ ኣየር ከም ሩዝ፣ ዓይኒዓተርን ዓተርን ዝበሉ ዘራእቲን ከም ሰሰግን ኣወሱዳን ዝበሉ ቅመማትን ንምፍራይ ምቹው እዩ። & This climate is ideal for growing crops like rice, wheat, chickpeas, and spices such as basil and black seed.

Table IV
Sample Evaluation Dataset Snippet from the Educational Domain with Sentence-level Alignment

Morphological Synthesizer for Ge'ez Language: Addressing Morphological Complexity and Resource Limitations

Gebrearegawi Gebremariam[†], Hailay Teklehaymanot^{*},
Gebregewergs Mezgebe[†]

[†]Axum University, Institute of Technology, Department of IT, Ethiopia

^{*}L3S Research Center, Leibniz University Hannover, Germany
{gideygeb,gemezgebe}@aku.edu.et,teklehaymanot@l3s.de

Abstract

Ge'ez is an ancient Semitic language renowned for its unique alphabet. It serves as the script for numerous languages, including Tigrinya and Amharic, and played a pivotal role in Ethiopia's cultural and religious development during the Aksumite kingdom era. Ge'ez remains significant as a liturgical language in Ethiopia and Eritrea, with much of the national identity documentation recorded in Ge'ez. These written materials are invaluable primary sources for studying Ethiopian and Eritrean philosophy, creativity, knowledge, and civilization. Ge'ez is a complex morphological structure with rich inflectional and derivational morphology, and no usable NLP has been developed and published until now due to the scarcity of annotated linguistic data, corpora, labeled datasets, and lexicons. Therefore, we proposed a rule-based Ge'ez morphological synthesis to generate surface words from root words according to the morphological structures of the language. Consequently, we proposed an automatic morphological synthesizer for Ge'ez using TLM. We used 1,102 sample verbs, representing all verb morphological structures, to test and evaluate the system. Finally, we get a performance of 97.4%. This result outperforms the baseline model, suggesting that other scholars build a comprehensive system considering morphological variations of the language.

Keywords: Ge'ez, NLP, morphology, morphological synthesizer, rule-based

1. Introduction

Language is one of the most important aspects of our lives, as it allows us to preserve information and pass it on orally or in writing from generation to generation (Allen, 1995).

Ge'ez is an ancient Semitic language with a unique alphabet ("አፄ በ፣ ገ፣ ደ") (Adege and Manie, 2017; Siferew, 2013). This language played a pivotal role in Ethiopia and Eritrea's cultural and religious development during the Aksumite Kingdom era. Its rich literary tradition and influence in spreading Christianity across the region are notable. Although no longer spoken colloquially after the thirteenth century, Ge'ez remains significant as a liturgical language for various religious groups. Scholars and linguists are drawn to Ge'ez for its insights into the historical evolution of Semitic languages and their connections to languages such as Hebrew, Arabic, and the modern Ethiopian and Eritrean language (Dillmann and Bezold, 2003; Desta, 2010; Abate, 2014).

Besides being the liturgical language for various religious groups in Ethiopia and Eritrea, Ge'ez remains a significant writing language for religious, historical books, and literature in the history of Ethiopia (Belcher, 2012; Scelta and Quezzaire-Belle, 2001). These written resources can be primary sources for studying Ethiopian and Eritrean philosophy, creativity, knowledge, and civilization. (Abate, 2014).

Hence, preserving the Ge'ez language becomes imperative to safeguarding Ethiopia and Eritrea's cultural and historical heritage. As the language deeply intertwined with religious practices and literature, its preservation ensures the continuity of traditions and identities across generations. Besides, preserving the Ge'ez language is crucial for maintaining religious practices and literature traditions, honoring linguistic diversity and identity, contributing to the understanding of Semitic languages' evolution, and fostering cultural pride and continuity across generations in Ethiopia and Eritrea (Desta, 2010).

However, research for this language has only started recently, and no usable technology has been developed and published until now for the Ge'ez because little consideration has been given to the language, even though it is that important. Due to this, Ge'ez is still a low-resource and endangered language (Eiselen and Gaustad, 2023; Haroutunian, 2022). In documenting endangered languages or reconstructing historical languages, understanding their morphological structure is essential for accurately representing and preserving the linguistic systems (Bisang et al., 2006). For morphologically rich languages such as Ge'ez, it is essential to develop a system that can generate all surface word forms from root words because this can serve as an input for many other NLP systems, including IR systems, spelling and grammar checking, text prediction, dictionary development,

POS tagging, machine translation, conversational AI, and other AI-based systems. But, it is difficult to develop AI-based systems especially for low-resourced languages such as Ge'ez, etc (Eiselen and Gaustad, 2023; Haroutunian, 2022; Gasser, 2012; Saranya, 2008; Scelta and Quezzaire-Belle, 2001; Sunil et al., 2012; Wintner, 2014).

For example, consider the search results in Table 1 to evaluate the limitation of the IR system in Ge'ez word variation.

Queries	Verb Form	Results
ገጠኝ/reTene/	Perfective	9
ይገጠኝ/yrTn/	Indicative	0
ይገጠኝ/yrTn/	Subjective	0
ገጠኝ/rTin/	Noun	1,480

Table 1: Ge'ez queries and their results from the Google search engine

As shown in Table 1, the results obtained in each query are different, even though the queries are related and generated from the verb 'ገጠኝ/reTene/'. In this case, the query should be given in all variants of the word forms; if not, the system will fail to retrieve the related information. However, it is inconvenient to search for all variant words (Hailay, 2013). To improve the efficiency of IR systems, it is important to create a strong relationship between the stems and their variant word forms. Thus, it is important to develop a morphological synthesizer of Ge'ez and integrate it with the IR systems to get an effective IR system.

Therefore, we proposed a rule-based Ge'ez morphological synthesizer that can play a crucial role in generating surface words from the root words according to the morphological structures of the language. This study is the first attempt to develop morphological synthesizers for the Ge'ez language, although morphological synthesizers for other languages have been developed and are available for wider usage, as stated below in the related works section. As a result, our work has made the following fundamental contributions to the scientific community:

- i. We designed an algorithm based on the language's morphological rules to illustrate generating TAM and PNG features. We tried to create surface words from the lexicons. The generator uses Ge'ez Unicode alphabets without transliterating to Latin alphabets. This makes it easy to use, especially for Ge'ez learners and researchers.
- ii. We prepared the first publicly available datasets for Ge'ez morphological synthesizers. Another researcher can use it.
- iii. Our system gives Amharic and English meanings for the perfect verb form. Therefore,

this can initiate the development of the following higher Ge'ez-Amharic, Ge'ez-Tigrinya or Ge'ez-other languages dictionary projects.

2. Related Works

One of the most popular research areas in NLP is the study of morphological synthesizers. Several research projects have been conducted in this area for various international languages using different approaches (Abeshu, 2013; Koskenniemi, 1983). Let us look at some related works.

ENGLEX was developed to generate and recognize English words using TLM in PC-KIMMO. It has three essential components, including a set of phonological (or orthographic) rules, lexicons (stems and affixes), and grammar components of the word. The generator accepts lexical forms such as **spy** + **s** as input and returns the surface word spies. The online source code is available here¹.

Jabalín was developed for both analyzing and generating Arabic verb forms using Python. They created a lexicon of 15,453 entries. This was designed using a rule-based approach called root-pattern morphology. The morphological generator accepts verb lemmas to produce inflected word forms and achieved an accuracy of 99.52% for correct words (González Martínez et al., 2013).

Using a paradigm-based approach, the Morphological Analyzer and Synthesizer for Malayalam Verbs was also developed by (Saranya, 2008). This helps in creating an English-Malayalam machine translation system.

Pymorphy2 was developed for the morphological analysis and generation of Russian and Ukrainian languages (Korobov, 2015). The system used large and efficiently encoded lexicons built from Open-Corpora and LanguageTool data. A set of linguistically motivated rules was developed to enable morphological analysis and the generation of out-of-vocabulary words observed in real-world documents.

TelMore was developed by (Ganapathiraju and Levin, 2006) to handle the morphological generation of nouns and verbs in Telugu. The prototype was designed based on finite-state automata. TelMore accepts the infinitive form for the verb types and generates the present, past, and future tenses, affirmative, negative, imperative, and prohibitive forms for all genders and numbers. In addition, (Dokkara et al., 2017) also developed a morphological generator for this language. Its computational model was developed based on finite-state techniques. The system was evaluated for a total

¹<http://downloads.sil.org/legacy/pc-kimmo/engl20b5.zip>

of 503 verbs. Of these verbs, 418 words were correct, and 85 words were incorrect.

(Goyal and Lehal, 2008) developed the morphological analyzer and generator for Hindi using the paradigm approach. This system has been developed as part of the machine translation system from Hindi to Punjabi. (Gasser, 2012) developed a system that generates words for Amharic, Oromo, and Tigrinya words from the given root and affixes. This has been developed based on the concept of finite-state technology. The system produced 96% accurate results (Gasser, 2012).

A morphological synthesizer for Amharic was developed by (Lisanu, 2002) using combinations of rule-based and artificial neural network approaches. However, his study was limited to Amharic perfect verb forms. Some of the generated word forms could be more meaningful. Also, this model used a transliteration of the Amharic script into Latin before any synthesis was done. The system does not allow generation for other roots that are not registered in its database. On the other hand, words are generated as output by giving the root and suffix as inputs. This may limit the number of words the model can produce compared to the words developed by the language experts. (Lisanu, 2002).

(Abeshu, 2013) developed an automatic morphological synthesizer for Afan Oromoo using a combination of CV-based and TLM-based approaches and achieved a performance of 96.28 % for verbs and 97.46% for nouns. The study indicated that developing a full-fledged automatic synthesizer for Afan Oromoo using rule-based approaches can yield an outstanding result. And it is easy to extend the system to other parts of speech with minimal effort.

The morphological synthesizers reviewed overhead are specific to their corresponding language and cannot handle Ge'ez's morphological characteristics because Ge'ez differs from these languages. To our knowledge, no research has been conducted to develop an automatic morphological generator for the Ge'ez language. Thus, we planned to create a morphological synthesizer model that can generate the derivational and inflectional morphology of Ge'ez language verbs.

3. Ge'ez Morphology

Ge'ez language has a complex morphological structure because a single word can appear in many different forms and convey different meanings by adding affixes or changing the phonological patterns of the word (Adege and Mannie, 2017). In particular, verbs have a more complex structure than other POSs in Ge'ez. Thus, Ge'ez verbs are categorized into six principal classes in their forms labeled as perfective, indicative, infinitive, subjunctive, jussive, and gerundive verb forms. Each verb

form has five stem classes, and each verb stem will inflect by adding affixes to create different word forms (Desta, 2010). Generally, there are three phases to creating variant word forms in Ge'ez, as defined in (Dillmann and Bezold, 2003). These are given below, as depicted in Figure 1:

Phase I: Stem formation

Phase II: TAM formation

Phase III: PNG formation

In Phase I, the declaration of word forms using the Tense-Mood as rows and the five stems as columns is done.

In Phase II, each surface verb form obtained from Phase I is further declared using the ten subjective pronouns by appending the subject marker suffix.

In Phase III, declarations of the word forms using the ten Object Marker Suffixes for each of the words obtained in Phase II will occur.

So, two rules for suffixing verbs govern the concatenation process of morphemes to produce the surface verb forms:

- Stem + subject-marker suffix = surface word (only with SMS)
- Stem + subject-marker suffix + object-indicator suffix = surface word (with both SMS and OMS)

Hence, we can have two verb forms, one with the only direct subject marker and the other with both subject marker and object marker suffixes, as indicated below:

$\Phi\text{-}\mathbf{t}\mathbf{A}$ (stem) + $\mathbf{h}\mathbf{m}$ (subject marker suffix) = $\Phi\text{-}\mathbf{t}\mathbf{A}\mathbf{h}\mathbf{m}$ - you killed. (Surface Form).

$\Phi\text{-}\mathbf{t}\mathbf{A}$ (stem) + $\mathbf{h}\mathbf{m}$ (object marker suffix) = $\Phi\text{-}\mathbf{t}\mathbf{A}\mathbf{h}\mathbf{m}$ - he killed you (Surface Form).

$\Phi\text{-}\mathbf{t}\mathbf{A}$ (stem) + $\mathbf{h}\mathbf{m}$ (SMS) + \mathbf{z} (OMS) = $\Phi\text{-}\mathbf{t}\mathbf{A}\mathbf{h}\mathbf{m}\mathbf{z}$ - you killed me (Surface Form).

In this case, the subject marker suffix /- $\mathbf{h}\mathbf{m}$ / points out that the subject is "you (2 ppm)," whereas the object marker /- \mathbf{z} / indicates the object "me." Hence, the verb / $\Phi\text{-}\mathbf{t}\mathbf{A}\mathbf{h}\mathbf{m}\mathbf{z}$ / indicates both the subject and the object of the verb. Hence, a single verb can be a sentence in Ge'ez because it has both subject and object indicator suffixes.

4. Methodology of the study

We have reviewed several books, research reports, journals, articles, and user manuals to grasp the morphological structure of Ge'ez verbs and to know the different techniques for designing morphological synthesizers. In addition, continuous discussions were conducted with Ge'ez experts to better understand the morphological structure of the language better and to get valuable ideas for the study.

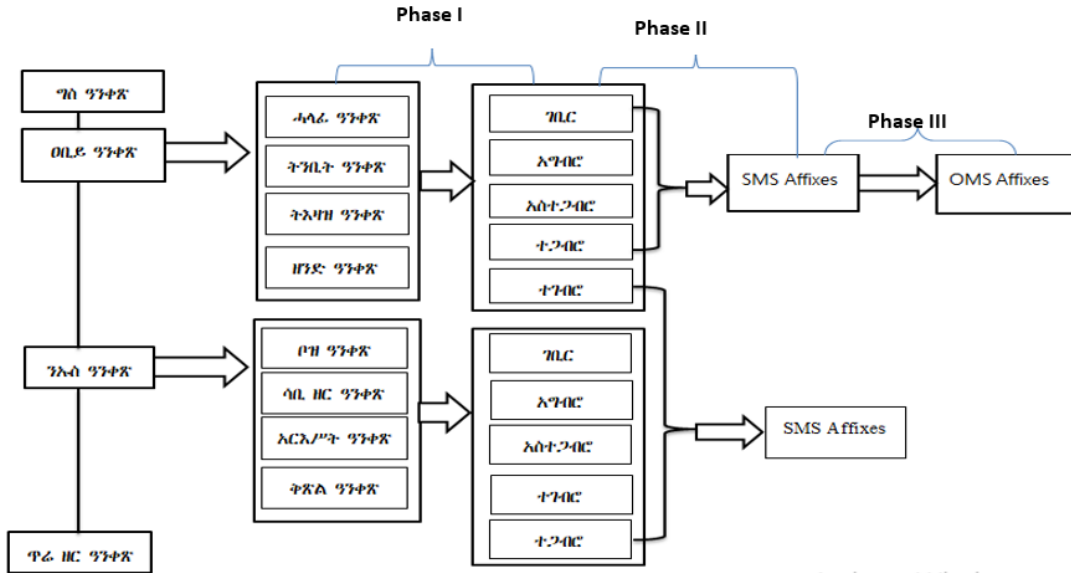


Figure 1: Phases of Ge'ez morphological word formation

4.1. Data Collection

Manually annotated data in lexicons helps test the morphological synthesizer. Since machine-readable dictionaries and word lists or an online corpus for Ge'ez were not available, the work of compiling the lexicons was started from scratch. Hence, we have compiled sample representative verbs that characterize all variations of verbs for testing and evaluating the systems's performance by consulting experts of the language. These verbs are collected from different books, like the Holy Bible, መጽሐፈ ግስ/Ge'ez Grammar Book/, and from Isanate siem (ልሳናተ ሴም) (Zeradawit, 2017). Therefore, the language lexicon prepared for this study consists of 1102 regular and irregular verbs. The affixes that can be concatenated with the verbs are also compiled into the lexicons.

4.2. Design

As defined by (Pulman et al., 1988), it is mandatory to consider at least the following basic design requirements to develop a morphological synthesizer of a language:

1. Lexicons:

Lexicon describes the list of all lemmas and all their forms. It is the heart of any natural language processing system, even though the format differs according to their needs. Consequently, the lexicons required for our study include stems, affixes, and Ge'ez alphabets. Let us see each of these lexicons in detail. i. **Stems:** In our study, the stem inputs are infinitive verb forms like ቀጥል/to kill/, ለዋር/to walk/, ሰጊድ/to Prostrate/, ፈቂድ/to allow/, ለዩው/to salivate/, etc. From these lexical inputs, the system generates inflected words for all genders and numbers by combining them with the corresponding affixes according to the set of rules

of the language. The reason why we want to use the infinitive verb form as input instead of the root word/ጥሬ ዘር/ is to remove the ambiguity that may be created when the prototype distinguishes the input's verb category.

ii. **Affixes:** As defined by (Abebe, 2010), the affixes carry different types of syntactic and semantic information, helping to construct various words. Affixes combine with the word stems to generate various words based on the set of rules. Here, Verbal-Stem-Marker Prefixes and Person-Marker Prefixes are combined first with the input stem to generate various word stems (Abebe, 2010). Then, SMS and OMS suffixes follow in sequence. For example, consider the formation of ይቆጥህ/He will kill you/ using TLM in Table 2.

As indicated in Table 2, for every stem to combine with affixes, an analyzer should investigate the type of stem and the affixes that can concatenate properly to create valid surface words. Hence, a set of rules was established to handle such requirements.

iii. **Ge'ez Alphabets:** As described by (Koskeniemi, 1983), both the lexical and surface-level words in the two-level model are strings extracted from the language alphabets. The lexical-level strings may contain some characters that may not occur on the surface-level strings. Accordingly, Ge'ez words are constructed by the meaningful concatenation of Ge'ez alphabets. The alphabets in the Ge'ez language include all the characters starting from *u*/he/ to *z*/fe/ and the four other complex-compound alphabets. All the alternations of characters in the lexical strings during surface word formations are retrieved from these alphabets. Implementing these alterations is handled based on the rules in the system prototype. The two-level rules are used here to specify the permis-

sible differences between lexical and surface word representations.

2. Morphotactics:

Morphotactics is the model or rule of morpheme ordering that explains which classes of morphemes can follow other courses of morphemes inside a word. Ge'ez verbs have their own rules for ordering the morphemes. The order of morphemes in the word formation of Ge'ez verbs is as follows:

[Prefix] + [Prefix Circumfixes] + [stem] + [Suffix Circumfixes] + [SMS] + [OMS]

3. Orthographic Rules:

Orthographic rules are the spelling rules that are used to model the changes that occur in a word when two morphemes combine. Therefore, a set of rules is essential in mapping input stems to surface word forms. These rules are designed based on the morphological nature of Ge'ez for each sequence of the word formation process. Ge'ez has its own spelling rules when morphemes are concatenated with each other. For example $\Phi\text{-}\mathbf{\Lambda}\text{-}\text{qetele}/+\text{-}\mathbf{h}/\text{ku}/$: $\Phi\text{-}\mathbf{\Lambda}\mathbf{h}\text{-}\text{qetelku}/$ (here, $\mathbf{\Lambda}/\text{e}/$ is changed to $/\text{h}/$ when $\{\Phi\text{-}\mathbf{\Lambda}\text{-}\text{qetele}/\}$ is added to the SMS $\{\mathbf{h}/\text{ku}/\}$).

By taking the above design requirements into account, we designed the general flow chart of the system as shown in Figure 2: As we see in the flow chart in Figure 2, the design of morphological synthesizer has the following components:

A. Stem Classifier: identifies the verb category of the stem. The classification is undertaken based on the number of heads and troops of verbs. This component also checks whether the verb stem is regular or not. Here, if the input verb contains one of the guttural alphabets (namely $\mathbf{u}/\text{he}/$, $\mathbf{h}/\text{He}/$, $\mathbf{\gamma}/\text{H}/$, $\mathbf{h}/\text{a}/$ and $\mathbf{o}/\text{A}/$ either at their beginning or middle positions) or semi-vowel alphabets (namely $\mathbf{f}/\text{ye}/$ and $\mathbf{w}/\text{we}/$) at any positions of the verb, it is irregular, else it is regular verb.

B. Stems Formation: This sub-component generates the various derived stems for the lexical input.

C. Signature Builder: lists the set of suffixes valid for each generated stem because every created stem has specific corresponding affixes to the stem during valid surface word formation. To establish a valid concatenation of the stems with affixes, a pattern matching mechanism is used, which is based on the notion of matching the stems with their valid affixes. For example, the word ' $\mathbf{\rho}\mathbf{\Phi}\mathbf{\Lambda}$ '/yqetl/ has a valid affix ' $\mathbf{\rho}$ '/wo/ to create a valid word form. But, this word cannot be combined with the affix ' $\mathbf{h}\mathbf{\rho}$ '/kwo/ because the combination of the word and the affix cannot create valid word forms.

D. Boundary Change Handler: This sub-component addresses the boundary patterns oc-

curing during the concatenation of stems and affixes based on the rules laid down on the knowledge base. These changes may be specific to every morpheme concatenation, even if these morphemes are in the same manner. Assimilation effects are occurring mostly on the boundary of the morphemes when the suffixes $\mathbf{h}/\text{ke}/$, $\mathbf{h}/\text{ku}/$, $\mathbf{h}/\text{ki}/$, $\mathbf{h}\mathbf{\gamma}/\text{kn}/$ or $\mathbf{h}\mathbf{o}/\text{kmu}/$ are added to the end of a verb that ends with either of the glottal alphabets, namely $\Phi/\text{QE}/$, $\mathbf{h}/\text{ke}/$, or $\mathbf{\gamma}/\text{Ge}/$ (Lambdin, 1978). For example, observe the concatenation of the morphemes $\mathbf{h}\mathbf{\rho}\mathbf{\gamma}$ with $\mathbf{h}\mathbf{o}$:

$\mathbf{h}\mathbf{\rho}\mathbf{\gamma} + \mathbf{h}\mathbf{o} \rightarrow \mathbf{h}\mathbf{\rho}\mathbf{\gamma}\mathbf{o}$ (the character $\mathbf{\gamma}$ in $\mathbf{h}\mathbf{\rho}\mathbf{\gamma}$ changes to $\mathbf{\gamma}$ and the character \mathbf{h} is omitted from the morpheme $\mathbf{h}\mathbf{o}$)

E. Synthesizer: This sub-component generates all possible surface word forms by concatenating the stem with the selected list of affixes using the TLM method of word generation. For example, consider the following Ge'ez word generation by TLM from Table 2:

Lexical Level	$\mathbf{\rho}$	Φ	$\mathbf{\gamma}$	$\mathbf{\Lambda}$	+	\mathbf{h}
Surface Level	$\mathbf{\rho}$	Φ	$\mathbf{\gamma}$	$\mathbf{\Lambda}$	0	\mathbf{h}

Table 2: Generation of surface words using TLM

The rows in Table 2 depict the two-level mappings carried out during the word formation process.

F. Surface Level: Lastly, the outputs of the synthesizer are produced.

Below is our concise algorithm for producing word forms based on input lexicons:

- ```

.....
1. Start
2. Input infinitive verb stem (verb stem)
3. Classify verb regularity using
 classifyVerbRegularity(verbstem)
4. If regular:
4.1 For each stem in generateStems(verb stem):
4.1.1 Select affixes with selectAffixes(stem)
4.1.2 Apply boundary changes
 with applyBoundaryChanges(stem)
4.1.3 Concatenate changed stems with affixes
4.1.4 Print output words
5. Else (if irregular):
5.1 For each stem in generateStems (verbstem):
5.1.1 Select affixes with selectAffixes(stem)
5.1.2 Apply boundary changes with
 applyBoundaryChanges(stem)
5.1.3 Concatenate changed stems with affixes
5.1.4 Print output words
6. End

```

## 5. Experimentation and Evaluation

### 5.1. Developmental Approach

Several approaches could have been applied to developing morphological generation systems for

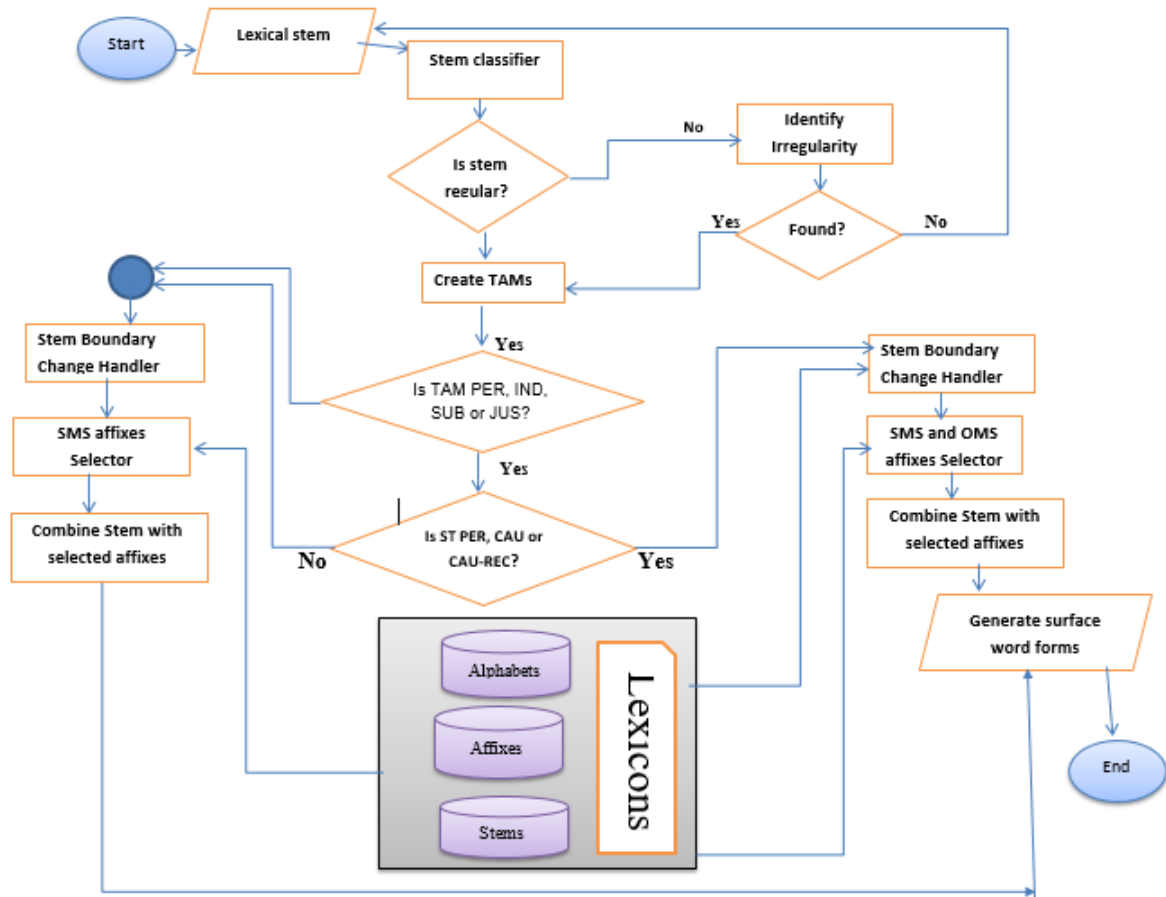


Figure 2: Flow Chart of Ge'ez morphological synthesizer

different languages. As discussed by (Kazakov and Manandhar, 2001), these approaches can be categorized as rule-based and corpus-based approaches. This study applied the rule-based approach called the Two-Level Model (TLM) of morphology to develop the prototype. TLM is used to handle the phonological and morphophonemic processes (including assimilation changes) involved in word formation (Gasser, 2011; Koskeniemi, 1984). Principally, we selected the TLM approach to map lexical entries to surface verb forms. We used the rule-based approach to develop the morphological synthesizer of the language because this approach has a faster development process with better accuracy, is more straightforward to twist, and is more accessible for formulating rules according to the language rules (Beesley and Karttunen; Shaalan et al., 2007). Moreover, the rule-based approach is practical for languages with fewer resources, such as Ge'ez, which suffers from the availability of corpora and the scarcity of data (Shaalan et al., 2010). Hence, we preferred the rule-based approach, in which a particular word is given as an input to the morphological synthesizer, and if that corresponding morpheme or root word is valid, then the system will produce surface word forms.

## 5.2. Testing Procedures

Systematic evaluation of the system is complex since no collected Ge'ez words are currently available for this purpose. So, to test the effectiveness of the system developed, we used the collected sample verbs. The testing procedures are as follows:

1. During the initial phase, we evaluated the system by inputting a test stem extracted from sample verbs in the lexicon, generating words, and comparing them with their expected word forms. This evaluation was conducted iteratively throughout the development of the morphological synthesizer to enhance its performance. Any errors identified during this testing, primarily related to missing rules, were rectified accordingly. Subsequent iterations of this test were conducted until satisfactory results were achieved.

2. Then, the finalized system's functionality was tested by entering sample verbs (including those with glottal or semivowel alphabets at different positions) selected by linguists.

## 5.3. Evaluation Procedures

Finally, we evaluated by taking regular and irregular verbs from the selected sample verbs. To evaluate the system, we used two options:

**1. Manual Evaluation** Using the error-counting approach, language experts manually evaluated the generated words to assess their accuracy and quality. The system accuracy is then calculated as the number of correctly generated words divided by the total number of words generated by the system multiplied by 100%.

**2. Automatic Evaluation** We evaluate system performance using predefined criteria and metrics without human intervention, a method akin to that described by (González Martínez et al., 2013). Subsequently, the accuracy attained from each experiment is calculated using the following formula:

$$\text{Accuracy} = \frac{\text{Correctly generated words}}{\text{total generated words}} * 100 \quad (1)$$

## 6. Experimental Results

The accuracy assessment of the developed system involved inputting sample datasets. 7,577 words were generated from regular verbs, and 19,290 words were generated from irregular verbs. Out of these, 668 errors were identified (8 from regular verbs and 661 from irregular verbs). The accuracy rates were: 99.6% for regular verbs and 96.6% for irregular verbs. Resulting in an overall average accuracy of 97.4%. This result Surpasses the baseline (Abeshu, 2013). The percentage of words with errors was 2.6%. This promising outcome supports further research on the language. The experimental results are found and referred in the Appendix section.

## 7. Discussion

The system consistently produces accurate words, albeit with occasional errors. As Appendix I details, irregular verbs perform less than regular verbs, primarily due to their inherent flexibility in word-formation processes. The predominance of irregular verbs in the evaluation dataset contributes to the observed decrease in accuracy. If a more significant proportion of regular verbs were included in the evaluation, the accuracy would be expected to surpass 97.4%, given the higher accuracy rate of 99.6% observed for words generated from regular verbs.

### 7.1. Factors Leading to High Performance.

Despite encountering some errors, the synthesizer demonstrates remarkably high performance. This achievement can be attributed to several factors:

**1. Creating correct stems** Correctly generated stems generate correct surface words if the boundary changes happening during stem and affix concatenations are handled correctly. If the root stems

are developed perfectly, then the words generated from these stems are correct. Hence, the performance achieved is high because most of the stems caused are right, and the boundary changes are handled correctly.

**2. Handling of rules when morphemes are concatenated with each other** Correct words are generated when stems and affixes are concatenated properly. For this reason, the selection of affixes for the given stem was handled properly. Therefore, handling the set of rules for word formation properly will generate valid words.

**3. Handling rules for irregular word formation** Ge'ez language has many irregular verbs. Irregular verbs are those that have a slight change in their morphological structure when compared to regular verbs. This is mostly happening due to the existence of one of the guttural alphabets, namely *u/he/*, *h/He/*, *ʾ/H/*, *h/a/* and *o/A/* either at their beginning or middle positions, or the existence of the semi-vowel alphabets, namely *ʾ/ye/* and *o/we/* at either position of the verbs. Irregular verbs have various rules to generate the correct word forms. These rules have slight differences from these regular word formation rules. Handling these rules of word formation gives you better accuracy. Accordingly, we have tried to handle the word formation rules as much as possible.

### 7.2. Error Analysis

Certain words are generated incorrectly. These errors can be attributed to the following factors:

**1. Errors caused due to exceptional characters existing in the verb** Some verbs have special characteristics, even though these verbs seem to have the same form as the head verb. For example, the system was designed to handle the verbs that end with the characters *ϕ/qe/*, *h/ke/*, and *ʾ/ge/* because it is assumed that these verbs have the same morphological characteristics as other verbs. However, this may not always be true if we consider the morphological structure of the verbs *ሠረቀ/šereqe/*, *ሐደገ/Hedege/*, and *ለሐቀ/leHeqe/*. These verbs have differences due to the existence of guttural or semi-vowel characters, or both as shown in Table 3.

| Verbs       | Differences observed |             |             |            |
|-------------|----------------------|-------------|-------------|------------|
|             | Indicative           | Subjective  | Jussive     | Infinitive |
| ሠረቀ/sereke/ | ይሠርቅ/yserk/          | ይሥርቅ/yserk/ | ይሥርቅ/yserq/ | ሠረቀ/seriq/ |
| ሐደገ/Hedege/ | የሐደግ/yeHedg/         | ይሐደግ/yHdg/  | ይሐደግ/yHdq/  | ሐደግ/Hedig/ |
| ለሐቀ/leHeqe/ | ይለሐቅ/yIHq/           | ይለሐቅ/yIHq/  | ይለሐቅ/yIHq/  | ለሐቅ/IHq/   |

Table 3: Errors caused by exceptional characters

As we see in table 3, the letters written in red color in the words make a difference in each word formation process even though these words are categorized in the same verb category.

**2. Errors generated during concatenation of exceptional words with affixes**

Some of the generated words seem to be correct both grammatically and semantically, but they

are not correct words. For example, when the morphemes ከረም/kerem/ and ነ/ne/ are concatenated, they produce the word ከረምነ/keremne/ which is the correct word. In the same way, when the morphemes አመን/amen/ + ነ/ne/, it gives አመንነ/ amenne/. However, አመን/ amenne/ is not the correct word. The correct word is አመነ/amene/. So, these words have different forms even though they belong to the same POS and number.

### 3.Errors caused due to morphological richness and varied nature of the language

This type of error occurs when testing with verbs that seems to have the same structure as other verbs in nature. But their actual output shows different word forms. For example, when we take the verbs ወለደ/welede/ and ወቀሰ/weqese/, we assume that these verbs have the same structure during the design of the prototype. But these verbs have differences in their actual word formation structure.

**4. Errors caused by missing some rules** The formation of the different word forms has a set of rules. Missing any of these rules generates invalid word forms. The incorrect words in Table 4 are generated because some rules and their correct forms are missing.

| ሙራሌ ግስ (pronoun)     | Incorrectly generated words | Correct words |
|----------------------|-----------------------------|---------------|
| ወ-አ-ቱ (He)           | ተከብ/tekebbe/                | ተከበ/tekebe/   |
| ይ-አ-ቲ (She)          | ተከብት/tekebbet/              | ተከት/tekebet/  |
| ወ-አ-ቶሎ (They-male)   | ተከብቱ/tekebbu/               | ተከቱ/tekebu/   |
| ወ-አ-ቶን (They-female) | ተከብታ/tekebba/               | ተከታ/tekeba/   |

Table 4: Errors caused due to missing rules

## 8. Conclusion and Future Work

The study opted for the rule-based TLM approach for developing an automatic morphological synthesizer due to its simplicity, suitability, and effectiveness, especially for languages with limited corpora availability. A set of rules was meticulously designed based on expert knowledge of the language's morphological structure, forming the foundation for algorithm development from scratch to handle word formation processes. Despite the thoroughness of the morphological synthesis rules, some inaccuracies persisted in word generation, mainly stemming from the formation of invalid stems, notably for irregular verbs containing guttural and semi-vowel alphabets. Nevertheless, the prototype synthesizer exhibited promising performance, with an overall accuracy of 97.4%, indicating encouraging prospects for further research in Ge'ez linguistics. Feedback from linguists involved in the system evaluation underscored the importance of developing a comprehensive system version to enhance Ge'ez's usage and preservation within society. Recommendations were made for future researchers to address and rectify errors limiting the study's performance and to

advance toward a fully functional system. Challenges encountered during the study included: A lack of Ge'ez linguistic experts. Absence of standardized references and dictionaries. Scarcity of compiled Ge'ez language lexicons. Furthermore, the complexity and agglutinative nature of Ge'ez morphology posed additional hurdles, contributing to its extensive vocabulary.

## List of Acronyms

|               |                                    |
|---------------|------------------------------------|
| EOTC.....     | Ethiopian Orthodox Tewahido Church |
| TLM.....      | Two-Level Morphology               |
| NLP.....      | Natural Language Processing        |
| PNGs.....     | Persons, Numbers and Genders       |
| POS.....      | Parts Of Speech                    |
| TAM.....      | Tense-Aspect-Mood                  |
| SMS.....      | Subject Marker Suffixes            |
| OMS.....      | Object Marker Suffixes             |
| IR.....       | Information Retrieval              |
| CV.....       | Consonant-Vowel                    |
| PER.....      | Perfective                         |
| IND.....      | Indicative                         |
| SUB.....      | Subjective                         |
| JUS.....      | Jussive                            |
| ST.....       | Stem Type                          |
| CAU.....      | Causative                          |
| CAU-REC... .. | Causative-Reciprocal               |
| XML.....      | Extensible Markup Language         |

## References

- Yitayal Abate. 2014. Morphological analysis of ge'ez verbs using memory based learning.
- A. Abebe. 2010. Automatic morphological synthesizer for afaan oromoo. A thesis Submitted to School of Graduate Studies of addis ababa University in Partial fulfillment for degree masters of Science in Computer Science.
- Abebe Abeshu. 2013. Analysis of rule based approach for afan oromo automatic morphological synthesizer. *Science, Technology and Arts Research Journal*, 2(4):94–97.
- Abebe Belay Adege and Yibeltal Chanie Mannie. 2017. *Designing a Stemmer for Ge'ez Text Using Rule based Approach*. LAP LAMBERT Academic Publishing.
- James Allen. 1995. *Natural language understanding*. Benjamin-Cummings Publishing Co., Inc.
- Kenneth R Beesley and Lauri Karttunen. Finite-state morphology: Xerox tools and techniques. *CSLI, Stanford*, pages 359–375.

- Wendy Laura Belcher. 2012. *Abyssinia's Samuel Johnson: Ethiopian Thought in the Making of an English Author*. OUP USA.
- Walter Bisang, Hans Henrich Hock, Werner Winter, Jost Gippert, Nikolaus P Himmelmann, and Ulrike Mosel. 2006. *Essentials of language documentation*. Mouton de Gruyter.
- Berihu Weldegiorgis Desta. 2010. *Design and Implementation of Automatic Morphological Analyzer for Ge'ez Verbs*. Ph.D. thesis, Addis Ababa University.
- August Dillmann and Carl Bezold. 2003. *Ethiopic grammar*. Wipf and Stock Publishers.
- Sasi Raja Sekhar Dokkara, Suresh Varma Penu-mathsa, and Somayajulu G Sripada. 2017. Verb morphological generator for telugu. *Indian Journal of Science and Technology*, 10:13.
- Roald Eiselen and Tanja Gaustad. 2023. Deep learning and low-resource languages: How much data is enough? a case study of three linguistically distinct south african languages. In *Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 42–53.
- Fitsum Gaim, Wonsuk Yang, and Jong C Park. 2022. Geezswitch: Language identification in typologically related low-resourced east african languages. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6578–6584.
- Madhavi Ganapathiraju and Lori Levin. 2006. Telmore: Morphological generator for telugu nouns and verbs. In *Proceedings of the Second International Conference on Digital Libraries*.
- Michael Gasser. 2011. Hornmorpho: a system for morphological processing of amharic, oromo, and tigrinya. In *Conference on Human Language Technology for Development, Alexandria, Egypt*, pages 94–99.
- Michael Gasser. 2012. Hornmorpho 2.5 user's guide. *Indiana University, Indiana*.
- Alicia González Martínez, Susana López Hervás, Doaa Samy, Carlos G Arques, and Antonio Moreno Sandoval. 2013. Jabalín: a comprehensive computational model of modern standard arabic verbal morphology based on traditional arabic prosody. In *Systems and Frameworks for Computational Morphology: Third International Workshop, SFCM 2013, Berlin, Germany, September 6, 2013 Proceedings 3*, pages 35–52. Springer.
- Alicia González Martínez, Susana López Hervás, Doaa Samy, Carlos G. Arques, and Antonio Moreno Sandoval. 2013. Jabalín: A comprehensive computational model of modern standard arabic verbal morphology based on traditional arabic prosody. In *Systems and Frameworks for Computational Morphology*, pages 35–52, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Vishal Goyal and Gurpreet Singh Lehal. 2008. Hindi morphological analyzer and generator. In *2008 First International Conference on Emerging Trends in Engineering and Technology*, pages 1156–1159. IEEE.
- B. Hailay. 2013. Design and development of tigrigna search engine. A thesis Submitted to School of Graduate Studies of addis ababa University in Partial fulfillment for the Degree of Master of Science in Computer Science.
- Levon Haroutunian. 2022. Ethical considerations for low-resourced machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 44–54.
- Dimitar Kazakov and Suresh Manandhar. 2001. Unsupervised learning of word segmentation rules with genetic algorithms and inductive logic programming. *Machine Learning*, 43:121–162.
- Mikhail Korobov. 2015. Morphological analyzer and generator for russian and ukrainian languages. In *Analysis of Images, Social Networks and Texts: 4th International Conference, AIST 2015, Yekaterinburg, Russia, April 9–11, 2015, Revised Selected Papers 4*, pages 320–332. Springer.
- Kimmo Koskenniemi. 1983. *Two-level morphology: A general computational model for word-form recognition and production*. University of Helsinki. Department of General Linguistics.
- Kimmo Koskenniemi. 1984. A general computational model for word-form recognition and production. In *10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*. The Association for Computational Linguistics.
- Thomas O. Lambdin. 1978. *Introduction to Classical Ethiopic (Ge'ez)*. Harvard Semitic Studies - HSS 24.
- K Lisanu. 2002. *Design and development of automatic morphological synthesizer for Amharic perfective verb forms*. Ph.D. thesis, Master's thesis, school of Information Studies for Africa, Addis Ababa.

- Stephen G Pulman, Graham J RUSSELL, Graeme D Ritchie, and Alan W Black. 1988. Computational morphology of english.
- SK Saranya. 2008. Morphological analyzer for malayalam verbs. *Unpublished M. Tech Thesis, Amrita School of Engineering, Coimbatore.*
- Gabriella F Scelta and Pilar Quezzaire-Belle. 2001. The comparative origin and usage of the ge'ez writing system of ethiopia. *Unpublished manuscript, Boston University, Boston. Retrieved July, 25:2009.*
- Khaled Shaalan, Azza Abdel Monem, and Ahmed Rafea. 2007. Arabic morphological generation from interlingua: A rule-based approach. In *Intelligent Information Processing III: IFIP TC12 International Conference on Intelligent Information Processing (IIP 2006), September 20–23, Adelaide, Australia 3*, pages 441–451. Springer.
- Khaled Shaalan et al. 2010. Rule-based approach in arabic natural language processing. *The International Journal on Information and Communication Technologies (IJICT)*, 3(3):11–19.
- Muluken Andualem Siferew. 2013. *Comparative classification of Ge'ez verbs in the three traditional schools of the Ethiopian Orthodox Church*, volume 17 of *Semitica et Semitohamitica Berolinensia*. Shaker Verlag, Aachen.
- R Sunil, Nimtha Manohar, V Jayan, and KG Sulochana. 2012. Morphological analysis and synthesis of verbs in malayalam. *ICTAM-2012*.
- Shuly Wintner. 2014. Morphological processing of semitic languages. In *Natural language processing of Semitic languages*, pages 43–66. Springer.
- A. Zeradawit. 2017. ልሳናተ ሴም, 1st edition. ትንሳኤ ግተምያ ድርጅት, Addis Ababa, Ethiopia.

## Appendix I

| No.          | Verb Input   | Verb Form | Number of words Generated | Number of correctly Generated words | Number of wrongly Generated words | Accuracy     |
|--------------|--------------|-----------|---------------------------|-------------------------------------|-----------------------------------|--------------|
| 1.           | ፈቀደ/feqedel  | Regular   | 1269                      | 1269                                | 0                                 | 100%         |
| 2.           | አመካ/amenel   | Irregular | 590                       | 563                                 | 27                                | 95.4%        |
| 3.           | ሠረቀ/šereqel  | Irregular | 1262                      | 1262                                | 0                                 | 100%         |
| 4.           | ከደነ/kedene/  | Regular   | 1260                      | 1233                                | 27                                | 97.9%        |
| 5.           | ስስከ/sebeke/  | Irregular | 1262                      | 1262                                | 0                                 | 100%         |
| 6.           | ሐደነ/Hedege/  | Irregular | 1262                      | 1162                                | 100                               | 92.0%        |
| 7.           | መሐለ/meHele/  | Irregular | 580                       | 547                                 | 33                                | 94.3%        |
| 8.           | ቀነየ/qeneyel  | Irregular | 580                       | 580                                 | 0                                 | 100%         |
| 9.           | አበየ/abeyel   | Irregular | 580                       | 490                                 | 90                                | 84.4%        |
| 10.          | ጠወየ/Teweyel  | Irregular | 580                       | 570                                 | 10                                | 98.2%        |
| 11.          | ጠመየ/TeAme/   | Irregular | 1162                      | 1162                                | 0                                 | 100%         |
| 12.          | ሐዳየ/Hetseyel | Irregular | 580                       | 490                                 | 90                                | 84.4%        |
| 13.          | ከበጠ/zebeTe/  | Regular   | 1262                      | 1262                                | 0                                 | 100%         |
| 14.          | ሐመመ/Hememel  | Irregular | 580                       | 580                                 | 0                                 | 100%         |
| 15.          | ወላደ/welede/  | Irregular | 1262                      | 1262                                | 0                                 | 100%         |
| 16.          | ሐረደ/Herede/  | Irregular | 580                       | 580                                 | 0                                 | 100%         |
| 17.          | ሐለየ/Heleye/  | Irregular | 580                       | 490                                 | 90                                | 84.5%        |
| 18.          | ፈደየ/fedeyel  | Irregular | 580                       | 580                                 | 0                                 | 100%         |
| 19.          | ከወወ/kewewel  | Irregular | 580                       | 576                                 | 4                                 | 99.3%        |
| 20.          | ተለወ/telewel  | Irregular | 580                       | 580                                 | 0                                 | 100%         |
| 21.          | ከበበ/kebebe/  | Regular   | 1262                      | 1258                                | 4                                 | 99.7%        |
| 22.          | ሐተተ/Hetete/  | Irregular | 580                       | 576                                 | 4                                 | 99.3%        |
| 23.          | ወጠነ/weTenel  | Irregular | 580                       | 551                                 | 29                                | 95%          |
| 24.          | ረወየ/reweyel  | Irregular | 584                       | 584                                 | 0                                 | 100%         |
| 25.          | ለወለ/lewese/  | Irregular | 1262                      | 1262                                | 0                                 | 100%         |
| 26.          | ደደለ/tsedele/ | Regular   | 1262                      | 1262                                | 0                                 | 100%         |
| 27.          | ከረወ/zerewel  | Irregular | 580                       | 580                                 | 0                                 | 100%         |
| 28.          | ገደፈ/gedefel  | Regular   | 1262                      | 1262                                | 0                                 | 100%         |
| 29.          | ገረመ/gereme/  | Regular   | 1262                      | 1262                                | 0                                 | 100%         |
| 30.          | ወቀነ/weqese/  | Irregular | 1262                      | 1082                                | 180                               | 85.7%        |
| <b>Total</b> |              |           | <b>26,867</b>             | <b>26,179</b>                       | <b>688</b>                        |              |
|              |              |           |                           |                                     | <b>Average Accuracy</b>           | <b>97.4%</b> |

Results obtained by the experimentation of the system prototype

| Prefixes |     | Suffixes |    | Circumfixes |     |
|----------|-----|----------|----|-------------|-----|
| ኢ        | ያስተ | ኩ        | እ  | ሆሙ          | እ-እ |
| አ        | አስተ | ነ        | ኢ  | የሙ          | ን-እ |
| ያ        | እ   | ከ        | ትየ | ዋ           | ት-እ |
| ይ        | ን   | ኪ        | ትነ | ያ           | ት-ኡ |
| ት        | እት  | ከሙ       | ን  | ዎን          | ት-ኢ |
| ታ        | ንት  | ከን       | ከ  | ሆን          | ት-አ |
| ይት       | ተ   | አ        | ሃ  | የን          | ይ-እ |
| ትት       | ና   | ኡ        | ሁ  | ኒ           | ይ-ኡ |
| ታስተ      | የ   | አት       | ዎ  | ኮ           | ይ-አ |
| ነ        | ዘ   | አ        | የ  | ቱ           |     |
| አስ       | ለ   | የ        | ሙ  | ዎሙ          |     |
| ናስተ      |     | አ        |    |             |     |

Some of the Identified Ge'ez Affixes

# አርባሔ ግስ ዘልሳነ ግእዝ GE'EZ MORPHOLOGICAL SYNTHESIZER

Home Help

ለግብ አርባሔ ግስ (Enter Infinitive Verb Form):

ጎረቤ ለጎቱጎ ግስ (Select Verb TAM):

አርባሔ

ለጥፋጎ

| ግስ: ጠዕሙ ትርጉም: ቀመሰ to taste                     |        |         |         |        |        |         |         |         |          |         |          |
|------------------------------------------------|--------|---------|---------|--------|--------|---------|---------|---------|----------|---------|----------|
| ለጎቱጎ ግስ: ሐላላ/perfective/ ለጥፋጎ ግስ: ጎረቤ Cባታ ዕጢቶች |        |         |         |        |        |         |         |         |          |         |          |
| ሙራሐ ግስ                                         | ጥና ግስ  | ላተ      | ላነ      | ላከ     | ላከ     | ላከሙ     | ላከጎ     | ላተ      | ላተሙ      | ላተ      | ላተጎ      |
| ወላይ                                            | ጠዕሙ    | ጠዕሙኒ    | ጠዕሙነ    | ጠዕሙከ   | ጠዕሙከ   | ጠዕሙከሙ   | ጠዕሙከጎ   | ጠዕሞ     | ጠዕሞሙ     | ጠዕሞ     | ጠዕሞጎ     |
| ይላይ                                            | ጠዕሙት   | ጠዕሙትኒ   | ጠዕሙትነ   | ጠዕሙትከ  | ጠዕሙትከ  | ጠዕሙትከሙ  | ጠዕሙትከጎ  | ጠዕሞተ    | ጠዕሞተሙ    | ጠዕሞተ    | ጠዕሞተጎ    |
| ወላቶ                                            | ጠዕሞ    | ጠዕሞኒ    | ጠዕሞነ    | ጠዕሞከ   | ጠዕሞከ   | ጠዕሞከሙ   | ጠዕሞከጎ   | ጠዕሞታ    | ጠዕሞታሙ    | ጠዕሞታ    | ጠዕሞታጎ    |
| ወላቶጎ                                           | ጠዕሞግ   | ጠዕሞግኒ   | ጠዕሞግነ   | ጠዕሞግከ  | ጠዕሞግከ  | ጠዕሞግከሙ  | ጠዕሞግከጎ  | ጠዕሞግታ   | ጠዕሞግታሙ   | ጠዕሞግታ   | ጠዕሞግታጎ   |
| ላጎተ                                            | ጠዕሞከ   | ጠዕሞከኒ   | ጠዕሞከነ   |        |        |         |         | ጠዕሞከጎ   | ጠዕሞከጎሙ   | ጠዕሞከጎ   | ጠዕሞከጎጎ   |
| ላጎቲ                                            | ጠዕሞከ   | ጠዕሞከኒ   | ጠዕሞከነ   |        |        |         |         | ጠዕሞከጎተ  | ጠዕሞከጎተሙ  | ጠዕሞከጎተ  | ጠዕሞከጎተጎ  |
| ላጎትሙ                                           | ጠዕሞከሙ  | ጠዕሞከሙኒ  | ጠዕሞከሙነ  |        |        |         |         | ጠዕሞከጎታ  | ጠዕሞከጎታሙ  | ጠዕሞከጎታ  | ጠዕሞከጎታጎ  |
| ላጎትጎ                                           | ጠዕሞከጎ  | ጠዕሞከጎኒ  | ጠዕሞከጎነ  |        |        |         |         | ጠዕሞከጎታ  | ጠዕሞከጎታሙ  | ጠዕሞከጎታ  | ጠዕሞከጎታጎ  |
| ላነ                                             | ጠዕሞኮ   |         |         | ጠዕሞኮከ  | ጠዕሞኮከ  | ጠዕሞኮከሙ  | ጠዕሞኮከጎ  | ጠዕሞኮ    | ጠዕሞኮሙ    | ጠዕሞኮ    |          |
| ጎልነ                                            | ጠዕሞጎ   |         |         | ጠዕሞጎከ  | ጠዕሞጎከ  | ጠዕሞጎከሙ  | ጠዕሞጎከጎ  | ጠዕሞጎሁ   | ጠዕሞጎሁሙ   | ጠዕሞጎሁ   | ጠዕሞጎሁጎ   |
| <b>ለጥፋጎ ለጥፋጎ ግስ ለርባታ ዕጢቶች</b>                  |        |         |         |        |        |         |         |         |          |         |          |
| ወላይ                                            | ለጥፋጎ   | ለጥፋጎኒ   | ለጥፋጎነ   | ለጥፋጎከ  | ለጥፋጎከ  | ለጥፋጎከሙ  | ለጥፋጎከጎ  | ለጥፋጎ    | ለጥፋጎሙ    | ለጥፋጎ    | ለጥፋጎጎ    |
| ይላይ                                            | ለጥፋጎት  | ለጥፋጎትኒ  | ለጥፋጎትነ  | ለጥፋጎትከ | ለጥፋጎትከ | ለጥፋጎትከሙ | ለጥፋጎትከጎ | ለጥፋጎተ   | ለጥፋጎተሙ   | ለጥፋጎተ   | ለጥፋጎተጎ   |
| ወላቶ                                            | ለጥፋጎ   | ለጥፋጎኒ   | ለጥፋጎነ   | ለጥፋጎከ  | ለጥፋጎከ  | ለጥፋጎከሙ  | ለጥፋጎከጎ  | ለጥፋጎታ   | ለጥፋጎታሙ   | ለጥፋጎታ   | ለጥፋጎታጎ   |
| ወላቶጎ                                           | ለጥፋጎግ  | ለጥፋጎግኒ  | ለጥፋጎግነ  | ለጥፋጎግከ | ለጥፋጎግከ | ለጥፋጎግከሙ | ለጥፋጎግከጎ | ለጥፋጎግታ  | ለጥፋጎግታሙ  | ለጥፋጎግታ  | ለጥፋጎግታጎ  |
| ላጎተ                                            | ለጥፋጎከ  | ለጥፋጎከኒ  | ለጥፋጎከነ  |        |        |         |         | ለጥፋጎከጎ  | ለጥፋጎከጎሙ  | ለጥፋጎከጎ  | ለጥፋጎከጎጎ  |
| ላጎቲ                                            | ለጥፋጎከ  | ለጥፋጎከኒ  | ለጥፋጎከነ  |        |        |         |         | ለጥፋጎከጎተ | ለጥፋጎከጎተሙ | ለጥፋጎከጎተ | ለጥፋጎከጎተጎ |
| ላጎትሙ                                           | ለጥፋጎከሙ | ለጥፋጎከሙኒ | ለጥፋጎከሙነ |        |        |         |         | ለጥፋጎከጎታ | ለጥፋጎከጎታሙ | ለጥፋጎከጎታ | ለጥፋጎከጎታጎ |
| ላጎትጎ                                           | ለጥፋጎከጎ | ለጥፋጎከጎኒ | ለጥፋጎከጎነ |        |        |         |         | ለጥፋጎከጎታ | ለጥፋጎከጎታሙ | ለጥፋጎከጎታ | ለጥፋጎከጎታጎ |
| ላነ                                             | ለጥፋጎኮ  |         |         | ለጥፋጎኮከ | ለጥፋጎኮከ | ለጥፋጎኮከሙ | ለጥፋጎኮከጎ | ለጥፋጎኮ   | ለጥፋጎኮሙ   | ለጥፋጎኮ   |          |
| ጎልነ                                            | ለጥፋጎጎ  |         |         | ለጥፋጎጎከ | ለጥፋጎጎከ | ለጥፋጎጎከሙ | ለጥፋጎጎከጎ | ለጥፋጎጎሁ  | ለጥፋጎጎሁሙ  | ለጥፋጎጎሁ  | ለጥፋጎጎሁጎ  |
| <b>ለጥፋጎ ለጥፋጎ ግስ ለርባታ ዕጢቶች</b>                  |        |         |         |        |        |         |         |         |          |         |          |
| ወላይ                                            | ለጥፋጎ   |         |         |        |        |         |         | ለጥፋጎ    |          |         |          |
| ይላይ                                            | ለጥፋጎት  |         |         |        |        |         |         | ለጥፋጎተ   |          |         |          |
| ወላቶ                                            | ለጥፋጎ   |         |         |        |        |         |         | ለጥፋጎታ   |          |         |          |
| ወላቶጎ                                           | ለጥፋጎግ  |         |         |        |        |         |         | ለጥፋጎግታ  |          |         |          |
| ላጎተ                                            | ለጥፋጎከ  |         |         |        |        |         |         | ለጥፋጎከጎ  |          |         |          |
| ላጎቲ                                            | ለጥፋጎከ  |         |         |        |        |         |         | ለጥፋጎከጎተ |          |         |          |
| ላጎትሙ                                           | ለጥፋጎከሙ |         |         |        |        |         |         | ለጥፋጎከጎታ |          |         |          |
| ላጎትጎ                                           | ለጥፋጎከጎ |         |         |        |        |         |         | ለጥፋጎከጎታ |          |         |          |
| ላነ                                             | ለጥፋጎኮ  |         |         |        |        |         |         | ለጥፋጎኮ   |          |         |          |
| ጎልነ                                            | ለጥፋጎጎ  |         |         |        |        |         |         | ለጥፋጎጎሁ  |          |         |          |

Screenshot of Sample Generated words from the Synthesizer