Standardization or Normalization? The Role of Amharic Homophone Characters from an NLP Research Perspective

Tadesse Destaw Belay
tadesseit@gmail.com
CIC, IPN, Mexico City

Seid Muhie Yimam
seid.muhie.yimam@uni-hamburg.de
University of Hamburg

Getie Gelaye
getie.gelaye@uni-hamburg.de
University of Hamburg

Abinew Ali Ayele
abinewaliayele@gmail.com
University of Hamburg

Amharic is written in the Ge'ez alphabet called ፊደል *fidäl*. In Amharic writing, there are different characters with the same sound, and they are called **homophones**. The homophones comprise the *ha* sounds ሀ ⟨ha⟩, ሐ ⟨ḥa⟩, and ኀ ⟨ḫa⟩, the *a* sounds አ ⟨ʾa⟩ and ዐ ⟨ʿa⟩, the ⟨*sə́*⟩ sounds ሰ ⟨sä⟩ and ሠ ⟨śä⟩, and the ⟨*ṣə́*⟩ sounds ጸ ⟨ṣä⟩ and ፀ ⟨ṣ́⟩ with all including their seven consonant-vowel combinations [1].

There exist various transliteration systems of the Amharic homophone characters. One approach suggests **normalizing** homophones into a single representation to reduce redundancy, simplify language model development, and ease the teaching of Amharic as a foreign language [2, 3]. This means that instead of all seven *ha* sound characters ሀ ⟨ha⟩, ሃ ⟨hā⟩, ሐ ⟨ḥa⟩, ሓ ⟨ḥā⟩, ኀ ⟨ḫa⟩, ኃ ⟨ḫā⟩, and ኻ ⟨ḳā⟩, the character ሀ ⟨ha⟩ will be used. Instead of the four *a* sound characters, አ ⟨ʾa⟩, ኣ ⟨ʾā⟩, ዐ ⟨ʿa⟩, and ዓ ⟨ʿā⟩, the character አ ⟨ʾa⟩ will represent all the other characters.

On the other hand, there is a strong recommendation for **standardizing** the Amharic spelling. Amharic language experts proposed a strict application of homophone characters to write a word in order to convey the correct meaning. This would mean that Amharic should use a consistent spelling by preserving the existing homophone characters according to its ancestral language Ge'ez [4, 5, 6].

In this work, we assessed various Amharic NLP tasks, including POS tagging, Named Entity Recognition (NER), sentiment analysis, machine translation (MT), and emotion analysis using LLMs. Using available Amharic NLP datasets, we: (1) evaluated the performance of state-of-the-art large language models (LLMs) with and without normalization, (2) examined how LLMs represent Amharic homophone characters, (3) explored the adoption status of Amharic homophone characters in online text, and (4) highlighted key insights from the evaluations for standardization and/or normalization practices.

References

[1] We used the transliteration system for Amharic according to the journal *Aethiopica: International Journal of Ethiopian and Eritrean Studies.*

[2] Abate, Solomon Teferra, et al. "Parallel corpora for Bi-Lingual English-Ethiopian languages statistical machine translation." *Proceedings of the 27th International Conference on Computational Linguistics*. 2018.

[3] Menuta, F. "Over-differentiation in Amharic orthography and attitude towards reform." *The Ethiopian Journal of Social Sciences and Language Studies* (EJSSLS) 3.1 (2016): 3-32

[4] Cowley, Roger. "The standardization of Amharic spelling." *Journal of Ethiopian Studies* 5.2 (1967): 1-8.

[5] Gezmu, A. M., et al. "Manually annotated spelling error corpus for Amharic." arXiv preprint arXiv:2106.13521 (2021).

[6] Yacob, D. "Application of the double metaphone algorithm to Amharic orthography." arXiv preprint cs/0408052 (2004).