# Cognate Detection for Historical Language Reconstruction of Proto-Sabaean Languages: the Case of Ge'ez, Tigrinya, and Amharic

**Elleni Sisay Temesgen**[1], **Hellina Hailu Nigatu** [2], **Fitsum Assamnew Andargie**[1]
[1] Addis Ababa Institute of Technology, Ethiopia [2]University of California Berkeley, USA,
Correspondence: elleni.sisay@aait.edu.et, hellina_nigatu@berkeley.edu

## Abstract

Languages evolve, increasing the risk of losing ancestral languages. In this paper, we explore Historical Language Reconstruction (HLR) for Proto-Sabaean languages, starting with the identification of cognates–sets of words in different related languages that are derived from the same ancestral language. We (1) collect semantically related words in three Afro-Semitic languages from a three-way dictionary (2) work with linguists to identify cognates and reconstruct the proto-form of the cognates, (3) experiment with three automatic cognate detection methods and extract cognates from the semantically related words. We then experiment with in-context learning with GPT-4o to generate the proto-language from the cognates and use Sequence-to-Sequence (Seq2Seq) models for HLR.

## 1 Introduction

As languages evolve, words gain new meaning and lose old ones. This evolution can lead to a parent language splitting into multiple child languages; forming language families that share a common ancestral language. The ancestral language is referred to as a proto-language (Meloni et al., 2021). Historical Language Reconstruction (HLR) attempts to reconstruct proto-languages from patterns in *cognates*–words in child languages that have likely evolved from the same word in the proto-language.

To reconstruct proto-languages, we first have to identify cognates. A popular method in cognate set identification is to start with semantically related words–for instance, from multilingual dictionaries–and look for patterns of phonetic change across the languages. Once we have cognates, proto-word reconstruction can happen at different levels of the language family tree (e.g. Meloni et al., 2021; Kim et al., 2023). As we move up the language family tree, collecting cognate sets becomes difficult as we would need to collect parallel words from the multiple languages we are considering. This challenge becomes further pronounced for low-resourced languages–languages that have limited data and are expert-constrained (Nigatu et al., 2024).

In this paper, we use automated methods and human expertise to identify cognates from three low-resourced Afro-Semitic languages: Ge'ez, Tigrinya, and Amharic. For the three languages, we first prepare the Swadesh list[1]–which is a list of things and concepts used for historical comparative linguistics. We then worked with linguists to identify cognates from the Swadesh lists and to reconstruct the proto-form for each cognate. We used this data to test three different computational methods that we used to identify cognates from a large set of words collected from a three-way dictionary. Relying on its generative property, we experimented with GPT-4o in few-shot setting to generate proto-forms for the automatically identified cognates and experimented with Seq2Seq models for proto-language reconstruction.

**Contribution** To the best of our knowledge, there have been no attempts to collect cognate sets and reconstruct the proto-language for the Ethio-Semitic branch[2] of the Afro-Semitic language family. We contribute (1) a dataset of cognates for the three languages identified through human experts and automated methods(§4), (2) proto-forms constructed by human experts (§4.3), (3) benchmark results with Seq2Seq models trained for HLR (§5). We provide an analysis of the cognates and the performance of different models in HLR showing common patterns of errors (§5). We intend to release the data freely for research purposes; see our

[1] https://linguifex.com/wiki/Swadesh_list.

[2] While the literature refers to the language family branch as Ethio-Semitic, we acknowledge that some of the languages are also spoken in Eritrea. Hence, we refer to the proto-form of the languages as Proto-Sabaean referring to the languages' ancient Sabaean roots.



*Proceedings of the 31st International Conference on Computational Linguistics*, pages 7415–7422
January 19–24, 2025. 

data sharing statement in Appendix 10.5.

## 2 Related Works

**Cognate Set Identification** Traditionally, identifying cognates requires meticulous manual comparisons of lexicons across various concepts, demanding significant linguistic expertise. Early cognate identification approaches rely mostly on phonetic similarities and sound correspondence. With the growth of large-scale linguistic datasets and computational power, computational approaches for cognate detection have become popular: methods that rely on edit distance (e.g Levenshtein et al., 1966), clustering (e.g. List et al., 2017), and expectation-maximization techniques (e.g. MacSween and Caines, 2020) have become popular. By integrating deep learning and advanced computational models with traditional linguistic methods (Akavarapu and Bhattacharya, 2024) introduced transformer-based models that show better performance in capturing phonetic and contextual similarities across languages, outperforming traditional alignment techniques. Yet, challenges remain for ancient and low-resource languages that have little-to-no collected data and linguistic experts.

**Historical Language Reconstruction** Earlier HLR work like Bouchard-Côté et al. (2009) used the Monte Carlo variant of the Expectation-Maximization algorithm. List et al. (2022) introduced alignment and classification methods, where an alignment algorithm segments words into phonemes, which are then used to predict proto-phonemes with classifiers like Support Vector Machines (SVM). Ciobanu et al. (2020) approached the task as a Sequence-to-Sequence (Seq2Seq) problem using Conditional Random Fields (CRF) and N-gram features for Romance languages. Another significant contribution was made by Meloni et al. (2019) who employed a character-based GRU with attention mechanisms, allowing for automatic feature extraction and notable improvements in performance in Latin languages. Kim et al. (2023) used Seq2Seq Transformer model, incorporating language embeddings and positional embeddings to better handle long sequences and complex dependencies for Chinese languages. To the best of our knowledge, generative models have not been utilized for Historical Language Reconstruction.

## 3 Ethio-Semitic Languages

Ethio-Semitic languages are a subset within the Afro-Semitic family. Spoken primarily in Ethiopia and Eritrea, the branch includes seven languages[3] (see Figure 1). Proto-Sabaean serves as the reconstructed ancestor language for these Ethio-Semitic languages, representing their historical and linguistic origins (Huehnergard et al., 2013; Hetzron et al., 2018). In this work we foucs on three out of the seven languages: Ge'ez, Tigrinya, and Amharic. Appendix 10.2 gives details about each of the three languages in our study.

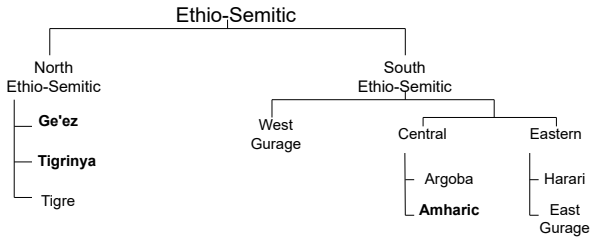


Figure 1: **Language family tree for the Ethio-Semitic branch of the Afro-Semitic language family.** We highlight in bold the languages included in our study.

Our languages of focus are low-resourced in that (1) there are limited digital resources in the languages, (2) there are constraints in accessing linguistic experts in the languages, and (3) there are computational and monetary constraints in developing and working on these languages (Nigatu et al., 2024). While there has been progress in these languages in text classification tasks (e.g. Neshir et al., 2020; Tela, 2020), machine translation (e.g. Ademtew and Birbo, 2024), and speech recognition (e.g. Abate et al., 2020), it is limited due to a lack of curated data, computational tools, and linguistic experts.

## 4 Data Collection

In this section, we will detail the three steps we took to collect cognates for HLR of Proto-Sabaean languages. Figure 2 shows our data curation steps.

### 4.1 Step 1: Semantically Equivalent Words

We started with the 100-concept Swadesh list in English and used dictionaries and human translation to curate Swadesh list in the three languages (see Figure 3). The next step was to collect larger data of semantically related words in the three languages.

[3]These languages include Ge'ez, Tigrinya, Tigre, Amharic, Argoba, Gurage (East Gurage and West Gurage ), and Harari.

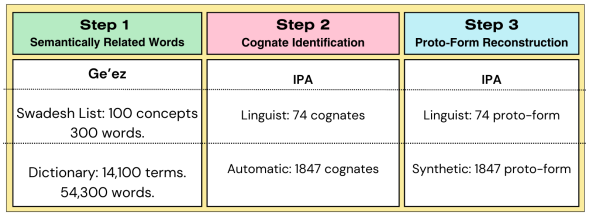


Figure 2: **Steps taken to create the dataset.** We show the number of words, cognates and proto-froms at each step.

| Swadesh Word | Ge'ez | Tigrinya | Amharic |
|---|---|---|---|
| All | ኩሉ<br>kʷilu | ኩሉ<br>kulu | ሁሉ<br>hulu |
| Egg | እንቁኦሆ<br>ʔəniqoqiho | እንቁላሊሕ<br>ʔiniqulaliħ | እንቁላል<br>inik'ulal |
| Eye | ዐይን<br>ʕəjin | ዓይኒ<br>ʕajini | ዓይን<br>ajin |
| Name | ስም<br>sim | ሽም<br>ʃim | ስም<br>sim |

Figure 3: **Samples from the Swadesh 100-word list in the three languages of study.** Below each word, we provide the IPA representation, which is what we used for the reconstruction and cognate identification.

We could not find a digital three-way dictionary in our languages of interest. Additionally, we could not rely on OCR methods as our task is sensitive to character errors and the current performance of such computational tools for these languages is low (Tonja et al., 2023). Hence, we had a three-way dictionary called Lsanat Sem, Adhana (1995) digitized by paid human experts.

**Results** This step resulted in 100 concepts translated to the three languages and a 54,300 word digitized version of a three-way dictionary; 14,100 words in each language. We then used the Epitran [4] library to convert the orthographic representation of the words to their IPA formats. This conversion enables us to represent the phonetic details of a language accurately, providing a phonetic transcription that reflects the actual sounds of spoken words.

### 4.2 Step 2: Cognate Identification

We worked with linguistics to identify cognates from the 100-concept Swadesh list described above. From the 100 concepts in the three languages, linguists identified 74 cognates. We used these cognates as a test set for automatic cognate set identification.

We used the LingPy[5] tool for our computational analysis. We experimented with three automatic cognate set detection methods: Turchin (Peter et al., 2010), Sound Class Algorithm (SCA) (List, 2010), and LexStat (List, 2012). Appendix 10.3 provides details of the three methods.

**Results** From the three automatic cognate identification methods, we found that SCA and LexStat performed better, accurately identifying 81% and 69% of the linguist-identified cognates (see Table 1). We applied the two methods on the words we collected from the three-way dictionary and retained sets of words that are identified by both SCA and LexStat as cognates. Since we did not have human expertise to review the full output, we took the two best performing methods as complementary ways of evaluation. From a dataset of 14,100 entries, we identified 1,847 cognates. Linguists then verified a 196 subset of the automatically identified cognates, finding errors in just 2% of the samples. This low error rate indicates that, even though the automatic cognate identification is not perfect, it can correctly identify a large set of cognates. Our final dataset from this step has 1847 cognates identified by automated methods and 74 cognates identified by linguistic experts.

| | TURCHINID | LEXSTATID | SCAID |
|---|---|---|---|
| Precision | 100 | 96 | 85 |
| Recall | 53 | 60 | 90 |
| F1 Score | 69 | 74 | 87 |
| Accuracy | 66 | 69 | 81 |

Table 1: **Results of automatic cognate identification techniques on linguist identified cognate sets.** We took two of the highest-scoring methods to automatically identify cognates from the three-way dictionary collected words.

### 4.3 Step 3: Proto-word Reconstruction

We first worked with linguists, who used the comparative method by (Anttila, 1989) to reconstruct the proto-forms for the 74 cognate sets they identified (see §4.2). Due to resource constraints to get linguistic expert reconstruction for 1847 cognate sets, we used GPT-4o to create the proto-forms. We first tested the performance of GPT-4o in constructing proto-forms for the 74 cognates from our human expert cognate identification (see §4.2). We then used GPT-4o in few-shot setting with 11 examples, selected based on linguists' judg-

[4] https://github.com/dmort27/epitran.

[5] https://lingpy.org/.

ments, to reconstruct the proto-form for the 1847 cognates from our automatic cognate identification. Appendix 10.4 shows the prompt structure we used. Since the Seq2Seq models are not trained on IPA representation for HLR task, the goal for the synthetic data is to inject the Seq2Seq knowledge into the models we wanted to test on the linguist-reconstructed data. We discuss the limitation with our approach in §7.

| Linguists | GPT-4o | Patterns | Explanation |
|---|---|---|---|
| bilaʕ | bəliʕə | Vowel Addition | Extra vowels (i, ə) added. |
| kəbd | kəris | Vowel & Consonant Substitution | *ə* → *i*, d → s. |
| ʕəbij | ʕabij | Vowel Substitution | *ə* → *a*. |
| hatʂ'ir | ħatʂ'ur | Vowel & Consonant Substitution | *i* → *u*, h → ħ. |
| kid | ħid | Consonant Substitution | *k* → *ħ*. |
| qətil | qətilot | Suffix Addition | *ot* added at the end. |
| nəkis | nəsik | Reordering | Final consonants (k → k) reordered |

Figure 4: Examples of patterns from the GPT-4o and the linguist reconstructed proto-forms.

**Results** On the linguist reconstructed proto-forms for the 74 cognate pairs, we find that GPT-4o had an 85% accuracy (see Table 2). In Figure 4, we show examples of errors that GPT-4o made on the linguist-reconstructed proto-forms. We then used it to generate proto-from for the 1847 cognates.

## 5 Experiments with Seq2Seq Models

With our curated dataset, we experiment with Seq2Seq models in reconstructing proto-forms. We had two test sets: (1) we used the human reconstructed proto-forms as one test set and (2) we split the dataset with 1847 cognates and synthesized proto-forms into train, validate and test set at an 8:1:1 ratio. We experimented with two Seq2Seq models: mT5 (Xue et al., 2021) and AfriTeVa (Jude Ogundepo et al., 2022). Model training details can be found in Appendix 10.1.

**Results** Table 2 presents our results. We find that mT5 significantly outperforms AfriTeVa. This is contrary to prior work which found community-centered models outperform generic models for low-resourced languages (Nigatu et al., 2023). We hypothesize this could be because mT5 is trained on a wider range of language varieties and hence might more easily learn the patterns in IPA format

| Model | Synthetic | Linguist |
|---|---|---|
| AfriTeVa-base | 8.45 | 12.23 |
| mT5-base | **48.40** | 57.74 |
| GPT-4o | - | **85.0** |

Table 2: **Accuracy performance of the model on synthetic and linguist reconstructed test set.** We do not report results for GPT-4o on the synthetic dataset since we used the model for synthesizing the data.

from the small training data we provided. While both mT5 and AfriTeVa include Amharic in their pre-training, we are training and testing in IPA format hence the prior knowledge of the language may not be as relevant. Our hypothesis is supported by the errors in the AfriTeVa reconstructed proto-forms which do not include the special IPA characters as compared to the linguist reconstructed proto-forms (see Figure 5). Figure 5 also shows a few patterns in sound change. For instance, we see some words that have no change in all three languages that are reconstructed by both Seq2Seq models accurately. We also observe language-specific patterns: Tigrinya words add /i/ phoneme at the end when compared to the proto-form. Amharic words omit glottal sounds /ʕ/ and /ʔ/. We also observe a shift in the /t͡sʼ/ sound to /tʼ/ in Amharic words while Tigrinya and Ge'ez retain the sound.

| Pattern | Ge'ez | Tigrinya | Amharic | Linguist | mT5 | afriTeva |
|---|---|---|---|---|---|---|
| No change | dəm | dəm | dəm | dəm | dəm | dəm |
| '/i/' addition in Tigrinya | tʂ'ifir | tʂ'ifiri | tʼifir | tʂ'ifir | tʂ'ifir | ts'fr |
| | midir | midiri | midir | midir | midir | mdr |
| '/ʕ/' and '/ʔ/' deletion in Amharic | nəʕa | nəʕa | na | nəʕa | nəʕa | nəa |
| | ʔigir | ʔigiri | igir | ʔigir | ʔigir | r |
| '/tʂ'/' shift to '/tʼ/' in Amharic | tʂ'ifir | tʂ'ifiri | tʼifir | tʂ'ifir | tʂ'ifir | ts'fr |
| | liħitʂ' | liħitʂ'i | lit' | liħitʂ' | liħitʂ' | lts' |

Figure 5: **Examples of patterns from the two Seq2Seq models and the linguist reconstructed proto-forms.** We observe here that AfriTeVa struggles to produce the IPA symbols in the reconstructed proto-forms. We also observe a few patterns, for instance, the addition of /i/ phoneme in the Tigrinya words when compared to the proto-forms.

## 6 Discussion and Future Work

Using human and automated methods, our work serves as a good starting point for the cognate identification and historical language reconstruction of Proto-Sabaean languages. We hypothesize the high performance of GPT-4o is because (1) it has seen a lot of IPA representations before, even if it might

not see our dataset, (2) it is larger and has higher capacity than the other models. However, there are several avenues for future work. First, the synthetic data can not be used as a fully reliable proto-form of the ancestral language. Rather, it serves as a way to train the Seq2Seq models to recognize the task of predicting proto-forms in IPA representation. Hence, the first avenue for future work would be to have linguists reconstruct the proto-form for the 1847 cognates. This requires resources which we currently lack.

The performance of AfriTeVa indicates that the model struggles to predict the special IPA symbols. In this paper, we use IPA because (1) it is how linguists perform HLR manually and (2) it will allow us to scale to other languages like Arabic in the future. For future work, we hope to run training and prediction in the orthographic representation of the languages (i.e. the Ge'ez script). We hypothesize that since mT5 and AfriTeVa both have Amharic in their pre-training data, training on the small dataset we have in Ge'ez script might improve performance as compared to using IPA. Further, since all three languages are phonetic (i.e. the words are spelled as they sound), the Ge'ez script version of the cognates will still reflect the sound changes.

We only looked at three out of seven of the languages in the Ethio-Semitic branch. Future work can add data from the other languages to improve the accuracy of the proto-language reconstruction. This might especially be helpful in preserving some of the languages like Arggoba which are currently endangered[6]. Finally, in future work, we also hope to add data from different dialects of the languages which will also improve the accuracy of proto-language reconstruction.

## 7 Limitations

The first limitation of our work is resources: we did not have enough resources to get expert reconstruction of our full dataset for the proto-form reconstruction. As a result, part of the dataset used in this research is constructed using automated methods, which may introduce errors. These inaccuracies can affect the performance of the proposed model, as the quality of machine-constructed data can vary and potentially impact the accuracy and reliability of the reconstruction process. Our focus in this paper is not the reconstruction of the proto-forms but rather the collection and identification of cognates. We rely on the linguist-reconstructed proto-forms to test the model performance and report on patterns we observed on that test set. Additionally, we only had three of the seven languages from the Ethio-Semitic branch. For future work, we hope to include data from the other languages.

[6] https://www.ethnologue.com/language/agj/.

## 8 Ethics Statement

This research includes the use of synthetic data. We try to minimize potential negative impacts by (1) using only human-reconstructed proto-forms to make claims about patterns in language change, (2) not openly releasing the synthetic data but rather releasing it only for research purposes with the proper declarations. We hope this work will inspire support and collaboration for access to resources that would help in having expert reconstruction of proto-form for the full dataset. Additionally, language is a sensitive cultural artifact. To minimize the potential misrepresentation of languages, we have worked with native speakers to verify the words we extracted from the three-way dictionary. We find that there are dialect differences that would lead to different words. We acknowledge this limitation and hope to address it in future work by collecting data from the different dialects to ensure representation.

## References


Solomon Teferra Abate, Martha Yifiru Tachbelie, Michael Melese, Hafte Abera, Tewodros Abebe, Wondwossen Mulugeta, Yaregal Assabie, Million Meshesha, Solomon Afnafu, and Binyam Ephrem Seyoum. 2020. Large vocabulary read speech corpora for four Ethiopian languages: Amharic, Tigrigna, Oromo and Wolaytta. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4167–4171, Marseille, France. European Language Resources Association.

Henok Ademtew and Mikiyas Birbo. 2024. AGE: Amharic, Ge'ez and English parallel dataset. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 139–145, Bangkok, Thailand. Association for Computational Linguistics.

Zra Dawit Adhana. 1995. *Lsanat Sem*. Kidst Slase Menfesawi Kolej, addis ababa.

VSDS Akavarapu and Arnab Bhattacharya. 2024. Automated cognate detection as a supervised link prediction task with cognate transformer. *arXiv preprint arXiv:2402.02926*.

Raimo Anttila. 1989. Historical and comparative linguistics.

Alexandre Bouchard-Côté, Thomas L Griffiths, and Dan Klein. 2009. Improved reconstruction of protolanguage word forms. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics*, pages 65–73.

Alina Maria Ciobanu, Liviu P Dinu, and Laurentiu Zoicas. 2020. Automatic reconstruction of missing romanian cognates and unattested latin words. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3226–3231.

A. Dolgopolsky. 1964. Gipoteza drevnejšego rodstva jazykovych semej severnoj evrazii s verojatnostej točky zrenija [a probabilistic hypothesis concering the oldest relationships among the language families of northern eurasia].

EKI Archive. n.d. Rom2-ti document. https://arhiiv.eki.ee/wgrs/v2_2/rom2_ti.pdf. Accessed: 2024-07-30.

Robert Hetzron, Alan S Kaye, and Ghil'ad Zuckermann. 2018. Semitic languages. In *The World's Major Languages*, pages 568–576. Routledge.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv*.

John Huehnergard et al. 2013. *The Semitic Languages*. Routledge.

Odunayo Jude Ogundepo, Akintunde Oladipo, Mofetoluwa Adeyemi, Kelechi Ogueji, and Jimmy Lin. 2022. AfriTeVA: Extending ?small data? pretraining approaches to sequence-to-sequence models. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 126–135, Hybrid. Association for Computational Linguistics.

Young Min Kim, Kalvin Chang, Chenxuan Cui, and David Mortensen. 2023. Transformed protoform reconstruction. *arXiv preprint arXiv:2307.01896*.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Johann-Mattis List. 2010. Sca: Phonetic alignment based on sound classes. In *European Summer School in Logic, Language and Information*, pages 32–51. Springer.

Johann-Mattis List. 2012. LexStat: Automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 117–125, Avignon, France. Association for Computational Linguistics.

Johann-Mattis List, Simon J Greenhill, and Russell D Gray. 2017. The potential of automatic word comparison for historical linguistics. *PloS one*, 12(1):e0170046.

Johann-Mattis List, Ekaterina Vylomova, Robert Forkel, Nathan Hill, and Ryan Cotterell. 2022. The sigtyp 2022 shared task on the prediction of cognate reflexes. In *Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 52–62. Association for Computational Linguistics.

Roddy MacSween and Andrew Caines. 2020. An expectation maximisation algorithm for automated cognate detection. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 476–485, Online. Association for Computational Linguistics.

Carlo Meloni, Shauli Ravfogel, and Yoav Goldberg. 2019. Ab antiquo: Neural proto-language reconstruction. *arXiv preprint arXiv:1908.02477*.

Carlo Meloni, Shauli Ravfogel, and Yoav Goldberg. 2021. Ab antiquo: Neural proto-language reconstruction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4460–4473, Online. Association for Computational Linguistics.

Girma Neshir, Solomon Atnafu, and Andreas Rauber. 2020. Bert fine-tuning for amharic sentiment classification. In *Workshop RESOURCEFUL Co-Located with the Eighth Swedish Language Technology Conference (SLTC), Gothenburg, Sweden*, volume 25.

Hellina Nigatu, Atnafu Tonja, and Jugal Kalita. 2023. The less the merrier? investigating language representation in multilingual models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12572–12589, Singapore. Association for Computational Linguistics.

Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. 2024. The zeno's paradox of 'low-resource' languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17753–17774, Miami, Florida, USA. Association for Computational Linguistics.

Turchin Peter, Peiros Ilia, and Gell-Mann Murray. 2010. Analyzing genetic connections between languages by matching consonant classes. , (5 (48)):117–126.

Abrhalei Frezghi Tela. 2020. Sentiment analysis for low-resource language: The case of tigrinya. Master's thesis, Itä-Suomen yliopisto.

Atnafu Lambebo Tonja, Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Moges Ahmed Mehamed, Olga Kolesnikova, and Seid Muhie Yimam. 2023. Natural language processing in Ethiopian languages: Current state, challenges, and

| Parameter | Value |
|---|---|
| training_batch_size | 32 |
| eval_batch_size | 8 |
| epochs | 100 |
| learning_rate | 1e-4 |
| lora_rank | 4 |
| lora_dropout | 0.01 |
| lora_alpha | 32 |

Table 3: **Hyperparamters for training mT5 and AfriTeVa models.**

opportunities. In *Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 126–139, Dubrovnik, Croatia. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

## 9 Acknowledgment

We extend our heartfelt gratitude to the 6 Kilo Linguistic Department, particularly Dr. Desalegn Hagos and his colleague, for their invaluable support in building the human-expert dataset and providing insightful guidance and expertise that have been crucial to the progress and success of this research.

## 10 Appendix

### 10.1 Model Paramteres

In this section, we provide the parameter details for the models we trained. Due to GPU constraints, we used LORA(Hu et al., 2021) for a parameter-efficient fine-tuning. We used the same hyperparameters for both mT5 and AfriTeVa. We experimented with a few learning rates [1e-4, 1e-5, 3e-4] and settled on 1e-4. See Table 3 for hyperparameter values.

### 10.2 Languages of Study

**Ge'ez** is an ancient Semitic language currently used in Ethiopian and Eritrean Orthodox churches. It belongs to the North Ethio-Semitic branch of the Afro-Asiatic language family. Linguistically, Ge'ez is distinguished by its Ge'ez script an 'Abugida' writing system where each symbol represents a consonant-vowel combination. This script has also been adapted to write other Ethiopian and Eritrean languages like Amharic and Tigrinya (Huehnergard et al., 2013; Hetzron et al., 2018).

**Tigrinya** is one of the official languages of Ethiopia and Eritrea and is spoken by 9.7 million people[7] in total across the two countries and their diasporas. Tigrinya uses the Ge'ez Script. Within the Tigrinya alphabet, there are 35 consonants and 7 vowels in the writing system (EKI Archive, n.d.).

**Amharic** is one of the official languages of Ethiopia, spoken by over 33.7 million people as a first language and 25.1 million as a second language according to the Central Statistical Agency of Ethiopia. Amharic also uses the Ge'ez script with an alphabet containing 32 consonants and 7 vowels. Amharic has undergone profound changes in its phonetic character: the laryngals have been reduced to 'h' sound and the glottal stop is rare now.

### 10.3 Automatic Cognate Identification Methods

**Turchin (Peter et al., 2010)** (also called Consonant Class Matching(CCM) approach ( following Dolgopolsky (1964) early idea to assume that words with two matching consonant classes would likely be cognate)) was proposed by Peter et al. (2010). In this method, the consonants of the words are converted to one of 10 possible consonant classes. The idea of consonant classes (also called sound classes) was proposed by Dolgopolsky, who stated that certain sounds occur more frequently in a correspondence relation than others and could therefore be clustered into classes of high historical similarity. In the approach by (Peter et al., 2010), two words are judged to be cognate, if they match in their first two consonant classes.

**Sound Class Algorithm (SCA) (List, 2010)** uses a threshold-based clustering algorithm and employs distance scores derived from the Sound-Class Based Alignment (SCA) method. This method for pairwise and multiple alignment analyses uses expanded sound class models along with detailed scoring functions as its basis. In contrast to previous alignment algorithms, the SCA algorithm takes prosodic aspects of the words into account and is

[7] https://en.wikipedia.org/wiki/Tigrinya_language#cite_note-E27-1.

also capable of aligning within morpheme boundaries, if morpheme information is available in the input data.

**LexStat (List, 2012)** method is based on flat UPGMA clustering, but in contrast to the SCA method, it uses language-specific scoring schemes which are derived from a Monte-Carlo permutation of the data. This permutation, by which the word lists of all language pairs are shuffled in such a way that words denoting different meanings are aligned and scored, is used to derive a distribution of sound-correspondence frequencies under the assumption that both languages are not related. The permuted distribution is then compared with the attested distribution, and converted into a language-specific scoring scheme for all language pairs. Using this scoring scheme, the words in the data are aligned again, and distance scores are derived which are then used as the basis for the flat cluster algorithm.

### 10.4 GPT-4o Prompt

The demonstration examples used for proto-form reconstruction with GPT-4o in section 4.3 are prepared as follows:

- Sampled 11 sequences, each containing 'lang1', 'lang2', 'lang3', and their corresponding 'proto' forms.
- Constructed input-output pairs for each sequence as $(x_1, f(x_1)), (x_2, f(x_2)), \ldots, (x_m, f(x_m))$, where $x_i$ represents a sequence in 'lang1', 'lang2', and 'lang3', and $f(x_i)$ represents the corresponding 'proto' form.

### 10.5 Data Sharing Statement

The main goal of this paper is to create a dataset of cognates for the historical language reconstruction of Proto-Sabaean languages. As such we intend to release the data with proper licences. We intend to make the cognate set data and the linguist-reconstructed cognates freely available for research purposes. We hope this effort will initiate more work in these languages and garner support to collect better quality and quantity data for these languages. To avoid misuse and improper utilization (for instance, discouraging collection of human-verified data), we will release the synthetic data for research purposes upon request and with proper declarations. We want to emphasize the labor that linguists afford to historical language reconstruction should be rewarded and that we do not make claims synthetic data can replace their efforts. All data and code will be released at https://github.com/ellenites/HLR.