Comparing methods of orthographic conversion for Bàsàá, a language of Cameroon

Alexandra O'Neil, Daniel Swanson
Robert Pugh, Francis Tyers
Indiana University
Bloomington, IN USA

Emmanuel Ngué Um
University of Yaoundé
Yaoundé, Cameroon

Abstract

Orthographic standardization is a milestone in a language's documentation and the development of its resources. However, texts written in former orthographies remain relevant to the language's history and development and therefore must be converted to the standardized orthography. Ensuring a language has access to the orthographically standardized version of all of its recorded texts is important in the development of resources as it provides additional textual resources for training, supports contribution of authors using former writing systems, and provides information about the development of the language. This paper evaluates the performance of natural language processing methods, specifically Finite State Transducers and Long Short-term Memory networks, for the orthographical conversion of Bàsàá texts from the Protestant missionary orthography to the now-standard AGLC orthography, with the conclusion that LSTMs are somewhat more effective in the absence of explicit lexical information.

1 Introduction

Orthographic standardization is a process that many languages of the world have undergone throughout history and many are still undergoing. Although there are numerous benefits to the standardization of a language's writing system, it can also present challenges for language communities. These challenges include contention between speakers that are used to using different representations, discomfort from speakers that relate to their language solely as an oral language, addressal and mitigation of the impact of colonialism on the language and community, debate about how to best represent sounds in the language, and hesitance in adoption of the writing system by all speakers (Limerick, 2018).

As referenced in the set of potential challenges above, communities often have differing means of representing their language prior to the coordinated effort to implement a uniform system (Mosel, 2004). While one of the goals of orthographic standardization is to create a consistent medium that speakers can use to understand one another and communicate their own thoughts, texts and data written in formerly used orthographies remain relevant in both the history and development of the language. To preserve this information it is necessary to convert former orthographies to the new standard. Furthermore, it is preferable to begin this process shortly following the standardization of the system, as this increases opportunity to work with speakers that are knowledgeable in the previously used systems.

Additionally, conversion of former orthographies into the current standard is beneficial since some speakers may not be willing to switch to the new standard. For a period of time following the adoption of the new orthography, speakers may continue to use a variety of orthographies in their own writing, following whichever orthography they previously learned (Jahani, 1989). Some speakers may be compelled to continue to use a non-standard system due to an emotional attachment to an orthographic system. For example, reasons for maintaining an orthographic preference range from positive experience, such as associating a system with how one's grandparents taught them, to a reaction to a traumatic experience, such as psychologically and physically abusive school environments where one writing system was emphasized (Arndt, 2019). Regardless of a speaker's reason for continuing use of a different orthography, it is constructive to the community to ensure that users of the new orthography are still able to understand writings in other orthographies and develop a method to easily convert these texts (Person, 2009).

In this paper, we investigate and compare the usefulness of finite-state transducers (FSTs)

Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023), pages 97–105

and long short-term memory neural networks (LSTMs) for the to the task of converting a prior orthography for the Bàsàá language to the current standard, with the conclusion that LSTMs slightly outperform FSTs in the absence of lexical information.

2 Bàsàá

Bàsàá is a Bantu language spoken by approximately 300,000 speakers in Cameroon (Eberhard et al., 2022). While it has many characteristic features of a Bantu language, it is commonly perceived to have more syllable structure variation and flexibility in noun classes when compared with other Bantu languages, as discussed in Section 2.1. The history of Bàsàá also accounts for a variety of writing systems from different missionaries and different standardization efforts, as outlined in Section 2.2.

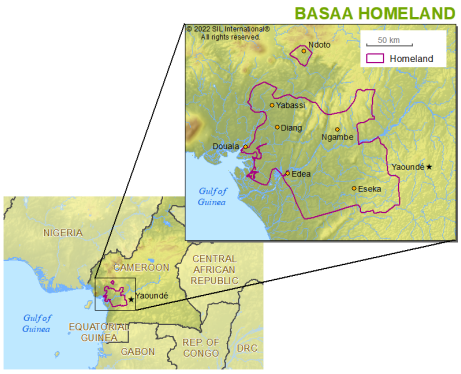

Figure 1: Map depicting the primary regions where Bàsàá is spoken: southern, central and littoral Cameroon. (Njock, 2019)

2.1 Linguistic Profile

A phonetic inventory of Bàsàá is laid out in Makasso and Lee (2015), which includes 7 phonemic vowels (see Figure 2) with short-long contrasts and 30 consonants (see Figure 3). Additionally, Bàsàá utilizes a high-low tone system. While it is a Bantu language, it atypically allows for closed syllable structure in addition to open syllable structure. Although it does have a noun class system, the surface distinctions between the classes are sometimes neutralized. Nouns are not required to start with a consonantal onset and verbs are not required to end in a vowel. These factors result in a higher diversity of permissible syllable structures in Bàsàá when compared with other Bantu languages (Hyman, 2003).

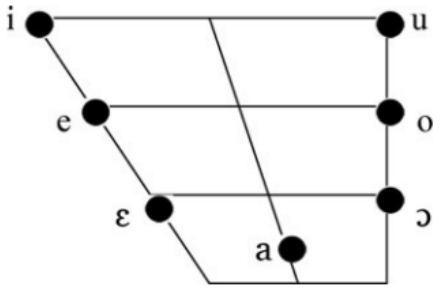

Figure 2: Vowel inventory in Bàsàá (Makasso and Lee, 2015). [ɛ] and [ɔ] are contrasted with [o] and [e] by diacritics in the missionary orthographies

2.2 Orthographic History

The orthographic history of Bàsàá contains multiple writing systems as English, German, and French colonialists who transcribed the language chose different methods of representation. The alphabets created for Bàsàá, and for many of the languages of Cameroon, were primarily influenced by the first languages of the transcriptionists, commonly French, English, or German, and often failed to properly mark important contrasts in the language (Bird, 2001). The two most prominent orthographies established prior to the current one are attributed to Protestant missionaries and Catholic missionaries and are referred to in this paper as the Protestant and Catholic orthographies, neither of which marked tone.

While the first attempt to implement a standard writing system, using an orthography developed for Western African language in Bamako (United Nations Educational and Organization, 1966), wasn't successful, a national committee was established to develop a writing system that would facilitate a pan-Cameroonian literacy in all the languages of the country. Central to this endeavor was the establishment of a system that could capture all of the contrasting sounds across Cameroonian languages. The inclusion of all contrasts would allow any literate speaker of a Cameroonian language to read and pronounce the words of a text in any of the languages of Cameroon, irrespective of comprehension (Hartell, 1993).

The national effort culminated in the establishment of the General Alphabet of the Cameroonian Languages (AGLC) by the National Committee for the Unification and Harmonisation of the Alphabets of Cameroon Languages in 1979 (Maurice Tadadjeu, 1979). The characters of this alpha-

	Bilabial	Alveolar	Post alveolar	Palatal	Velar	Labialized velar	Uvular	Glottal
Plosive	p	t			k	k^w g^w		
Affricate			tʃ dʒ					
Implosive	ɓ							
Prenasalized	^{m}b	^{n}d	^{n}dʒ		^{ŋ}g			
Nasal	m	n		ɲ	ŋ	$ŋ^w$		
Tap		ɾ̥ ɾ						
Fricative	ɸ β	s			x ɣ		χ	h ɦ
Approximant	w			j				
Lateral approximant		l						

Figure 3: Consonant inventory in Bàsàá (Makasso and Lee, 2015).

bet are predominately Latin, and thus similar to the English, German, and French alphabets, however the alphabet integrates symbols from the International Phonetic Alphabet to fully represent the phonetic inventory of Cameroonian languages. The symbols in the full AGLC are listed in Table 1.

Consonants	b,ɓ,c,d,ɗ,f,g,',h, j,k,l,m,n ŋ,p,q,r,s,t,v,w,ẅ,x,y,y̰,z
Vowels	a,ɑ,ɛ,e,ə,æ,ɤ,i,ɨ,o,ɔ,ø,œ,u,ʉ

Table 1: Alphabet of the Cameroonian Languages. The AGLC alphabet was designed to work as a unifying and intelligible alphabet for speakers of all Cameroonian language.

Bàsàá utilizes a subset of this system in accordance with the language's phonetic contrasts. Latin letters are used alongside ɓ, ŋ, ɛ, and ɔ with acute, grave, and circumflex accents to denoting tone. While this orthography is supported by the Academy of Languages of Cameroon, the former missionary orthographies are still used by some speakers and are generally used in earlier texts written in Bàsàá, necessitating a method of conversion of missionary orthographies to the AGLC standard.

Comparing the three writing systems, the first major discrepancy is found in tone marking and vowels. The Protestant and Catholic writing systems do not mark tone, but instead use acute, grave, and circumflex accents to mark different vowels. The Protestant system represents [e] by using an acute accent mark, é, while a [ɛ] is represented by a plain e. Likewise, the Protestant system uses the circumflexed ô, to mark [o], but a plain o in the orthography represents [ɔ]. On the other hand, the Catholic orthography marks [ɛ] and [ɔ] by è and ò, while the plain e and o represent [e] and [o]. However, in the AGLC orthography, acute, grave, and circumflex accents are used to represent tone and the ɛ and ɔ are already contrasted in the orthography by the addition of the symbols ɛ and ɔ. Tone is also marked on syllabic nasal consonants in the AGLC system.

The consonants in the orthography also require consideration in the conversion process. The [ɓ] and [ŋ] sounds are represented by b and ñ in the missionary orthographies, while the AGLC orthography distinguishes [b] from [ɓ] with the characters b and ɓ. Instead of ñ, the [ŋ] is represented by ŋ in the AGLC orthography. In addition to these consonants, the missonary orthographies transcribe the sound tʃ as tj, the AGLC system uses c. One sentence is presented in the three orthographies below to exemplify some of the differences in the orthographies.

AGLC:	Mɛ̀ yè lɛ mɛ ɓɔl nyɔɔ̄ nì màkòò.
Protestant:	Me yé le me bol nyoo ni makôô.
Catholic:	Mè ye lè mè bòl nyòò ni makoo.

3 Prior work

Negotiation between orthographies is a common issue in many languages, and as such a number of previous studies have explored techniques to aid in orthographic conversion and/or normalization. This work builds on existing research describing the Bàsàá language and the use of natural language processing for low-resource languages. This section outlines existing work on Bàsàá in Section 3.1, Finite State Transducers (FST) in Section 3.2, and Long Short-term Memory (LSTM) networks in Section 3.3. While research concerning FSTs and LSTM networks in low-resource settings is extensive, this section focuses on examples that

closely mirror the goals of this paper.

3.1 Research on Bàsàá

Existing work on Bàsàá has profiled the linguistic inventory of the language (Makasso and Lee, 2015; Hyman, 2003), generated dictionaries (Lemb and de Gastines, 1973), designed learning materials (Moreton et al., 1968), and, more recently, facilitated the development of resources that integrate computational and NLP methods with Bàsàá to enhance the resources available for documentation, such as the bilingual speech corpus developed to assist in automatic phonetic transcription (Hamlaoui et al., 2018).

Nikitin et al. (2022) approaches the task of orthography conversion in Bàsàá by using Bidirectional Encoder Representations from Transformers (BERT), which often performs well due to the large amount of resources and training that went into the model. However, for this task, BERT was only able to beat the baseline after extensive preprocessing of the text. The importance of extensive pre-processing the text and only marginally better performance than the baseline suggests that BERT is not well-suited to this task.

3.2 Finite State Transducers

Finite State Transducers (FST) work as a translator between a set of input strings and a set of output strings. In the case of language, the input string can utilize linguistic rules and produce an output that adheres to those rules. As the input of the FST relies on linguistic rules, the model often performs well in low-resource environments, as the models do not rely on large amounts of training data or computational resources. While FST models require some amount of linguistic or computational efforts to build, various tools have been created which automate various parts of the process to help alleviate these boundaries (Khanna et al., 2021), although the performance of FST models benefits greatly from the generation of detailed, language-specific rules.

In general, FSTs do not necessarily specify a unique mapping between input and output strings, which can cause problems for tasks like orthography conversion that generally need a single output. This can be addressed by adding additional rules to constrain the transducer. However, if determining appropriate rules is difficult and a corpus is available, it can also be addressed by adding weights, which are scores that can be applied either to a whole-word input-output pair or to a subword mapping. Then, for a given input, each output has a weight equal to the sum of all the applicable weights derived from the corpus, with only the form with the lowest total being output.

FSTs have been implemented in many low-resource settings, as well as for the application of orthographic conversion, transliteration, and text normalization. Washington et al. (2021) developed a transducer to assist in orthographic conversion and morphological analysis of Zapotec and found that even an incomplete transducer could yield positive results. Similar efforts use an FST to develop a morphological generator and analyzer while simultaneously addressing the issue of missing diacritics (Alkhairy et al., 2020), demonstrating the easy expansion of an FST to create more resources for a language. Manohar et al. (2022) extend the use of FSTs to text-to-speech (TTS) applications in low-resource settings, generating a model that converts between Malayalam phonemes and graphemes.

While the use of FSTs in low-resource settings is well-attested, the inclusion of tones has proven difficult. Ngué Um et al. (2022) built an FST for Ewondo, a Cameroonian language. In this study, the ambiguous nature of combined versus combining tone markings produces difficulty for the analyzer. While both the combined and combining accents can be analyzed by the FST, it results in errors in the morphological generation. An expansion of the FST for Bàsàá would also need to address this issue.

3.3 Long Short-Term Memory Networks

The LSTM is a recurrent neural network architecture that allows information about long-term dependencies to be incorporated, providing additional context for the generation of the output. LSTM networks have been applied to many deep learning tasks, such as machine translation, optical character recognition (OCR), and speech recognition.

LSTMs have been combined with OCR tasks to assist languages in digitization and orthographic normalization of historic texts. Azawi et al. (2013) found that LSTMs perform well for the conversion of German historic texts as they are able to handle unseen examples. Similarly, Simistira et al. (2015) found using LSTMs for OCR produces a lower character error rate than leading methods of

OCR for Greek polytonic script.

Additionally, LSTMs have recently gained popularity for their utility in TTS tasks, such as grapheme-to-phoneme conversion. Adriana (2019), Liu et al. (2018), and Behbahani et al. (2016) successfully implemented LSTM models for grapheme-to-phoneme conversion in Romanian, Mongolian, and Persian.

4 Methodology

This paper compares the accuracy of a Finite State Transducer (FST), a weighted FST, and a Long Short-term Memory (LSTM) model for the task of orthographic conversion of Bàsàá. Section 4.1 describes the data the models trained on, Section 4.2 describes the simple baseline metric used for comparison, Section 4.3 explains the use of an unweighted FST, Section 4.4 details the implementation of the weighted FST, and Section 4.5 outlines the LSTM model.

4.1 Data

The methods in this study use a text corpus comprised of 12,000 sentences in the Protestant orthography together with transliterations into the AGLC orthography. Of these sentences, 10,000 are used for training, 1,000 for validation, and 1,000 for testing. Pre-processing of the text consisted of lower-casing the characters.

4.2 Baseline

The baseline searches the target sentences for the most frequent translation of a source word and replaces the source word with that token. In the event that the source word does not appear in the data, the source word is output in its original form without conversion. The baseline here is a naïve approach to the problem, but is representative of the current lack of existing work on orthographic conversion in the language.

4.3 Unweighted FST

The unweighted FST consists of a set of character mappings compiled using the lexicon compiler Lexd (Swanson and Howell, 2021) which includes every pair of source and target characters found in the training data. Mapping each character individually and without context creates a large number of output forms, which we resolve by selecting a single output form at random. Additionally, we added four rules which restrict the output in cases where the phonological context is unambiguous. Specifically, that nasals will never have a tone diacritic in the AGLC orthography if they precede a vowel (this is 3 rules, one each for m, n, and ŋ), and that where the missionary orthography has tj the AGLC orthography will always have c (as opposed to converting the t and the j separately as tj).

4.4 Weighted FST

Where the unweighted FST treats every mapping as equally probable, the weighted FST sets a weight for each path which has been seen in the training data, with more frequent source-target pairs receiving lower (better) weights. Then, rather than selecting randomly, the output with the lowest weight is used.

4.5 Neural seq2seq model

Following previous work that has shown that character-based neural seq2seq architectures perform well for orthographic normalization and conversion (Rosca and Breuel, 2016; Orife, 2018), we trained an encoder-decoder model with global attention (Luong et al., 2015) to convert missionary orthographies into the AGLC orthography. Both the encoder and decoder are unidirectional Long short-term memory networks (Hochreiter and Schmidhuber, 1997) consisting of 2 layers of 1,000 hidden units each. We used the OpenNMT-py library (Klein et al., 2020) to train the model and generate predictions on the held-out datasets.

5 Results

We compare the four systems using word- and character-error rates (WER and CER). WER and CER are calculated automatically by comparing the outputs of each of the systems to the output of the target file. Following the presentation of WER and CER for each of the systems, we provide examples of the output from each model. Information on the differences in orthographical representation for the source and target texts can be found in Section 2.2.

5.1 Word- and Character-error Rates

Results of all systems apart from the unweighted FST were relatively similar, with baseline model performing better than both the FST systems and on par with the seq2seq model. The seq2seq model achieved the best character-error rate, while

System	CER	WER
Baseline	15.61	41.11
Unweighted FST	40.31	90.72
Weighted FST	18.06	55.10
LSTM	13.27	42.03

Table 2: Comparison between the 4 models.

the baseline shows a marginally better word-error rate. This can be explained by the fact that the baseline operates at the word level. Thus, a mistake results in selecting the wrong word form, which likely has multiple characters that are different than the correct word form. The seq2seq model, on the other hand, predicts at a character level, and may make only a single character error in a word, such as a missed tonal diacritic.

5.2 Error Analysis

To better understand the performance of these models, we present examples of outputs of the models for three different sentences and discuss which errors are common for each of the models. The examples are taken from the development set output for each of the models.

Source:	*malét a nhundus binan.*
Target:	màlêt à ǹhundus binan.
Baseline:	màlêt à nhundus bìnan.
Unweighted FST:	màlĕt à ŋhúńdus bi̯ŋan.
Weighted FST:	màlĕt à ŋhúńdus bi̯ŋan.
LSTM:	màlêt à ǹhundus binan.

Table 3

In the sentences in Table 3, we see that the FSTs have a tendency to overgenerate the letter ŋ when the source orthography has an n. Additionally, it shows that the LSTM network is successful in generating the tone of a syllabic nasal. While the baseline often can predict tone on a syllabic nasal, this token is not in the training data, so the baseline just outputs the original token.

The sentences in Table 4 show that the most systems are able to understand that the accents in the source orthography are not indicative of an accent on the target orthography. However, the unweighted FST tends towards adding accents even in the absence of accents in the target form. Overall, we see the LSTM perform well on the assignment of tone and characters in this example, with the baseline being almost perfect apart from the tone on the second token.

Source:	*nledek mut u nnéega bé.*
Target:	ǹlɛ̀dɛk mût u nnēegà ɓe.
Baseline:	ǹlɛ̀dɛk mùt u nnēegà ɓe.
Unweighted FST:	ǹlɛ̄dɛk mût ù ŋnêègâ ɓè.
Weighted FST:	ǹlɛ̀dɛk mùt ù nnēegà ɓe.
LSTM:	ǹlɛ̀dɛk mût u nnēegà ɓe.

Table 4

Source:	*kal nye le me ñke.*
Target:	kăl nyɛ lɛ mɛ̀ ŋ̀kɛ̂.
Baseline:	kăl nyɛ lɛ mɛ̀ ŋkɛ̀.
Unweighted FST:	kāl ńyɛ lɛ̆ mɛ̄ ŋ̀kɛ.
Weighted FST:	kâl nyɛ lɛ mɛ̀ ŋkɛ̀.
LSTM:	kal nyɛ̄ lɛ mɛ̀ ŋ̀kɛ̂.

Table 5

In the sentences in Table 5, although the LSTM comes close, it is not able to identify the rising tone in the first token and marks the tone of the second token as mid. The weighted and unweighted FST correctly predict the characters, but otherwise are very inconsistent with tone diacritics, although it is evident that the weighted FST outperforms the unweighted FST.

6 Conclusion

The results of this paper contribute to the discussion concerning the relative benefits of NLP methods versus more simplistic baselines in low-resource settings. The baseline outperforms the unweighted and weighted FSTs and LSTM network in regards to WER. As Bàsàá is a tone language, changes in the tone of one character can create minimal pairs and thus WER is a more realistic metric for evaluating the utility of a model. However, the baseline simply outputs the original word for any out-of-vocabulary (OOV) tokens, meaning that the performance of the baseline is strongly impacted by OOV tokens. While this has a minimal impact on this dataset, a dataset with more OOV tokens would perform worse.

While the baseline performs well on the Protestant orthography, it would likely perform even better on the Catholic orthography as it uses grave

accent marks instead of acute accent marks. Although the Catholic and AGLC orthography use the grave accents to mark different things, the presence of grave marks in the target is much more likely than acute accent marks, which only appear when deconstructing rising and falling tones on long vowels. The method of handling OOV tokens would therefore perform better for a source text written in the Catholic orthography, as the probability of coincidentally having an output that matches the input form is much higher when the source orthography uses grave accents.

In this paper, the unweighted and weighted FST were written using very minimal linguistic rules, which is evidenced in their relatively poor performance. The weighted FST greatly reduced the impact of the lack of detailed rules, but would clearly still benefit from their addition.

This paper presents a preliminary investigation of the application of FSTs and LSTM networks to the topic of orthographic conversion. While the simplistic baseline performs surprisingly well for this dataset, we believe that the comparable performance of the weighted FST and LSTM network is promising and necessitates further development of these models, specifically the inclusion of more linguistic rules for the weighted FST and augmentation of the training data for the LSTM network.

Limitations

This paper attempts to make a broader statements about the applicability of current NLP methods for text conversion by discussing the results of these models on Bàsàá. The case of Bàsàá is challenging as the representation of tones is difficult for many models. However, this study still benefits from the roman-based, alphabetic orthography of the language and the resources that are available to languages with a Latin-based, alphabetic orthography. Additionally, Bàsàá utilizes a transparent orthography that also facilitates automatic methods of conversion. Other results and challenges are likely to arise when applying these models to a language that utilizes a non-Latin-based, non-alphabetic, and/or opaque orthography.

As this project is intended to present a starting point for extended research on orthographic conversion, we have begun by providing an overall comparison and brief error analysis. However, we plan to implement a more systematic error analysis to guide future work. The current error analysis highlights some patterns that are observed in the data, but a more thorough review of the outputs will help in the development of the current system for Bàsàá.

Ethics Statement

The motivation of this work is to compare current NLP methods in a low-resource setting and discuss how the different systems might apply in different contexts based on the results, contributing overall to the discussion on how NLP methods can be used to benefit language communities and support the creation of more linguistic resources. While the hope is to support the language community, the integration of computational methods also poses the risk of language commodification and a dispossession of intellectual property of a community. This study is submitted with the belief that the current benefits associated with the application of this research outweigh this risk.

Acknowledgements

This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

References

STAN Adriana. 2019. Input encoding for sequence-to-sequence learning of romanian grapheme-to-phoneme conversion. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–6. IEEE.

Maha Alkhairy, Afshan Jafri, and David A Smith. 2020. Finite state machine pattern-root arabic morphological generator, analyzer and diacritizer. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3834–3841.

Jochen S Arndt. 2019. Engineered zuluness: Language, education, and ethnic identity in south africa, 1835–1990. *The Journal of the Middle East and Africa*, 10(3):211–235.

Mayce Al Azawi, Muhammad Zeshan Afzal, and Thomas M Breuel. 2013. Normalizing historical orthography for ocr historical documents using lstm. In *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing*, pages 80–85.

Yasser Mohseni Behbahani, Bagher Babaali, and Mussa Turdalyuly. 2016. Persian sentences to phoneme sequences conversion based on recurrent neural networks. *Open Computer Science*, 6(1):219–225.

Steven Bird. 2001. Orthography and identity in cameroon. *Written Language & Literacy*, 4(2):131–162.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2022. *Ethnologue: Languages of the World*, twenty-fifth edition. SIL International.

Fatima Hamlaoui, Emmanuel-Moselly Makasso, Markus Müller, Jonas Engelmann, Gilles Adda, Alex Waibel, and Sebastian Stüker. 2018. BUL-Basaa: A bilingual basaa-French speech corpus for the evaluation of language documentation tools. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Rhonda L. Hartell. 1993. Alphabets of africa. *SIL International*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Larry M Hyman. 2003. Basaá (a43). *The Bantu languages*, pages 257–282.

Carina Jahani. 1989. *Standardization and orthography in the Balochi language*. Ph.D. thesis, Acta Universitatis Upsaliensis.

Tanmai Khanna, Jonathan N Washington, Francis M Tyers, Sevilay Bayatlı, Daniel G Swanson, Tommi A Pirinen, Irene Tang, and Hèctor Alòs i Font. 2021. Recent advances in apertium, a free/open-source rule-based machine translation platform for low-resource languages. *Machine Translation*, 35(4):475–502.

Guillaume Klein, François Hernandez, Vincent Nguyen, and Jean Senellart. 2020. The opennmt neural machine translation toolkit: 2020 edition. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 102–109.

Pierre Lemb and François de Gastines. 1973. *Dictionnaire basaá-français*, avec un préface par meinrad hebga edition. Collège Libermann, Douala.

Nicholas Limerick. 2018. Kichwa or quichua? competing alphabets, political histories, and complicated reading in indigenous languages. *Comparative Education Review*, 62(1):103–124.

Zhinan Liu, Feilong Bao, and Guanglai Gao. 2018. Mongolian grapheme to phoneme conversion by using hybrid approach. In *Natural Language Processing and Chinese Computing: 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26–30, 2018, Proceedings, Part I 7*, pages 40–50. Springer.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Emmanuel-Moselly Makasso and Seunghun J. Lee. 2015. Basaá. *Journal of the International Phonetic Association*, 45(1):7179.

Kavya Manohar, AR Jayan, and Rajeev Rajan. 2022. Mlphon: A multifunctional grapheme-phoneme conversion tool using finite state transducers. *IEEE Access*, 10:97555–97575.

Etienne Sadembouo Maurice Tadadjeu. 1979. *Alphabet général des langues camerounaises / General Alphabet of Cameroon Languages*. Université de Yaoundé, SIL Internationale, University of Yaoundé, SIL International.

Rebecca L Moreton et al. 1968. Cameroon basaa. *Inspection copy available from Foreign Languages Program, Center for Applied Linguistics*.

Ulrike Mosel. 2004. Dictionary making in endangered speech communities. *Language documentation and description*, 2:39–54.

Emmanuel Ngué Um, Émilie Eliette, Caroline Ngo Tjomb Assembe, and Francis Morton Tyers. 2022. Developing a rule-based machine-translation system, ewondo–french–ewondo. *International Journal of Humanities and Arts Computing*, 16(2):166–181.

Ilya Nikitin, Brian O'Connor, and Anastasia Safonova. 2022. Tone prediction and orthographic conversion for basaa. *arXiv preprint arXiv:2210.06986*.

Pierre Emmanuel. Njock. 2019. *àsàa - French - English - German Dictionary.* Dallas: Webonary.org.

Iroro Orife. 2018. Attentive sequence-to-sequence learning for diacritic restoration of yorùbá language text. In *Interspeech*.

Kirk R Person. 2009. Heritage scripts, technical transcriptions, and practical orthographies: a middle path towards educational excellence and cultural preservation for thailands ethnic minority languages. In *Proceedings from the international conference on national language policy: Language diversity for national unity*.

Mihaela Rosca and Thomas Breuel. 2016. Sequence-to-sequence neural network models for transliteration. *arXiv preprint arXiv:1610.09565*.

Fotini Simistira, Adnan Ul-Hassan, Vassilis Papavassiliou, Basilis Gatos, Vassilis Katsouros, and Marcus Liwicki. 2015. Recognition of historical greek polytonic scripts using lstm networks. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 766–770. IEEE.

Daniel Swanson and Nick Howell. 2021. Lexd: A finite-state lexicon compiler for non-suffixational morphologies. In Mika Hämäläinen, Niko Partanen, and Khalid Alnajjar, editors, *Multilingual Facilitation*. University of Helsinki Library.

Scientific United Nations Educational and Cultural Organization. 1966. *Meeting of a group of experts for the unification of alphabets of national languages. Bamako, Mali, 28 February 5 March 1966. Final report*. United Nations Educational, Scientific and Cultural Organization, UNESCO.

Jonathan Washington, Felipe Lopez, and Brook Lillehaugen. 2021. Towards a morphological transducer and orthography converter for western tlacolula valley zapotec. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 185–193.